Statystyczna analiza danych

ukryte modele Markowa, zastosowania

Anna Gambin

Instytut Informatyki

Uniwersytet Warszawski

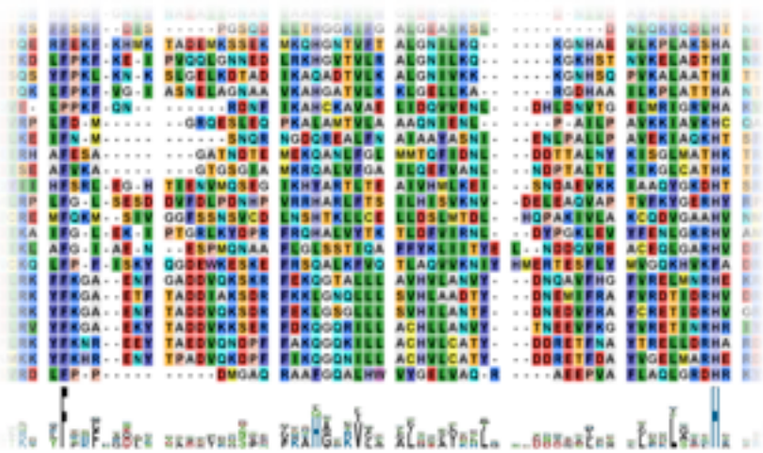
plan na dziś

Ukryte modele Markowa w praktyce

- modelowania rodzin białek
- multiuliniowienia

Pfam

- znajdowanie genów

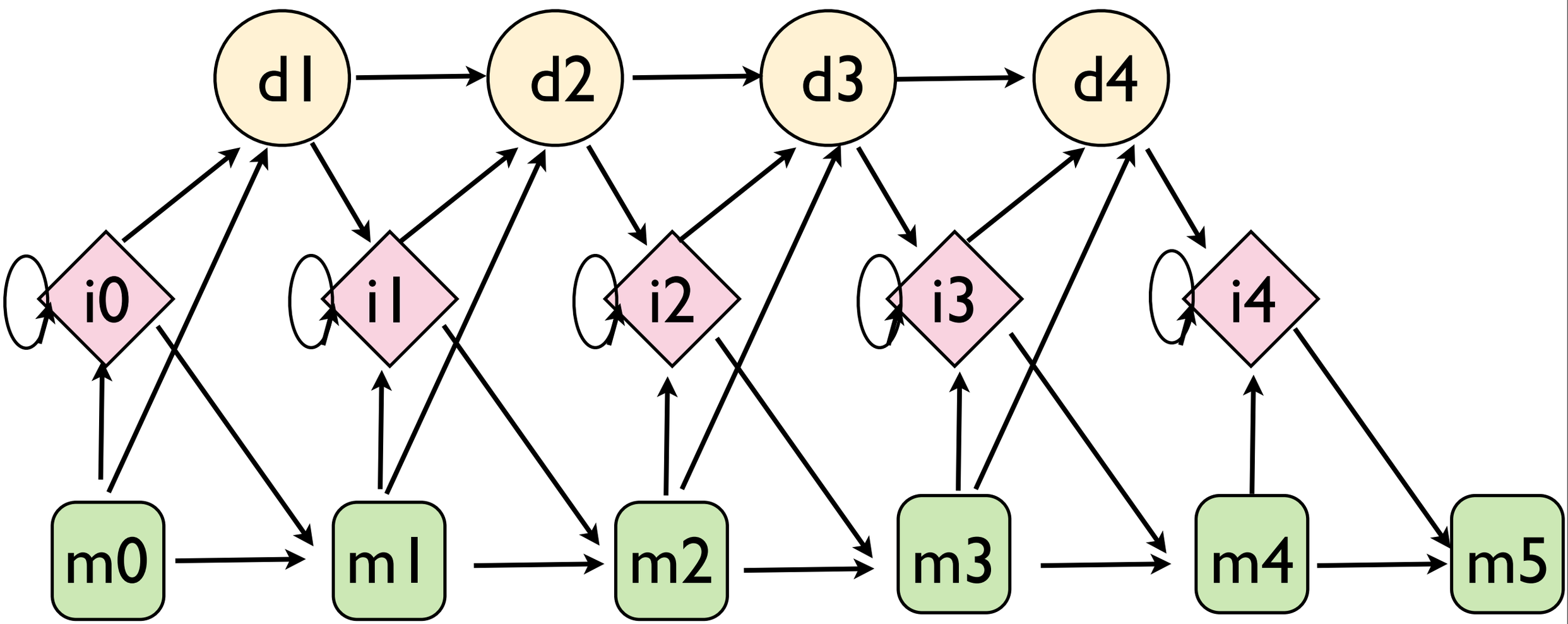


model rodziny białek

Cele:

- zbudowanie multiuliniowienia
- dla danej sekwencji identyfikacja rodziny

model rodziny białek



profilowy HMM

model rodziny białek

- długość sekwencji w przykładzie - 5
- stany: match, insert, delete
- m0 - stan początkowy, m5 - końcowy
- alfabet - 20 aminokwasów plus dziura δ

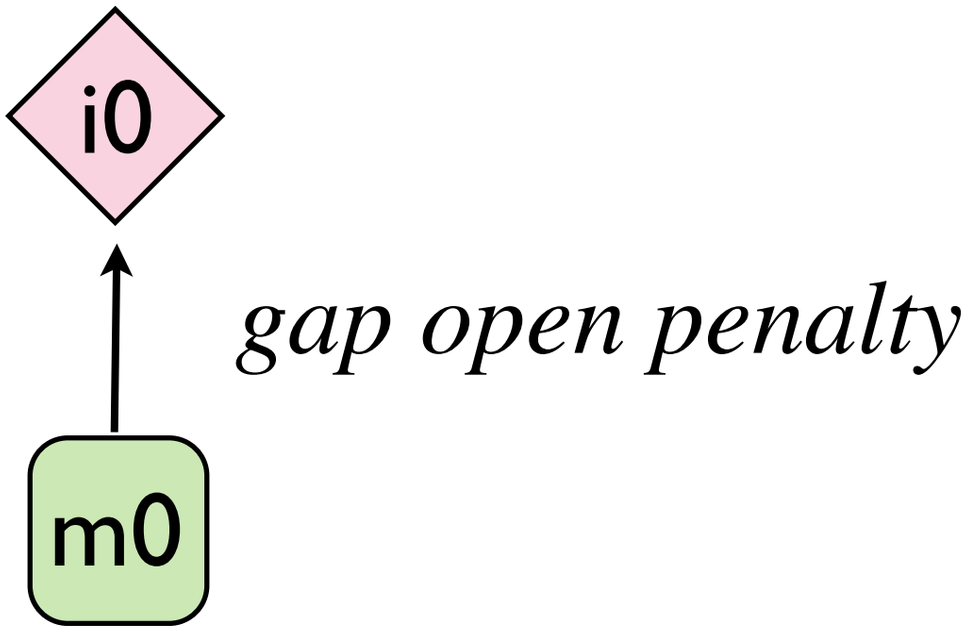
model rodziny białek

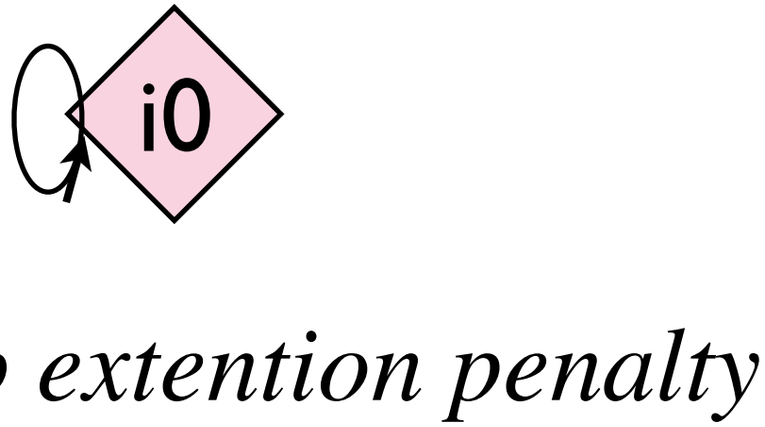
- każdy stan typu match i insert ma swój rozkład na aa, które emituje:
  - jeśli jest on jednostajny to emitujemy losowe sekwencje,
  - jeśli jednopunktowy to dokładnie jedną sekwencję,
  - jeśli „pomiędzy” to mamy pewną rodzinę sekwencji.
- stan typu delete emituje symbol dziury δ z prawdopodobieństwem 1

model rodziny białek

- dzięki modelowi hmm możemy zróżnicować podobieństwo różnych fragmentów sekwencji:
  - algorytmy dynamiczne i BLAST mają ustalony parametr na karanie przerwy oraz jedną tablicę substytucyjną dla całej sewencji.
  - hmm pozwala na zróżnicowanie tych parametrów w szczególności kara za przerwę może być w każdym miejscu inna.
- białka posiadają słabo i silnie konserwowane regiony (domenty funkcyjne).

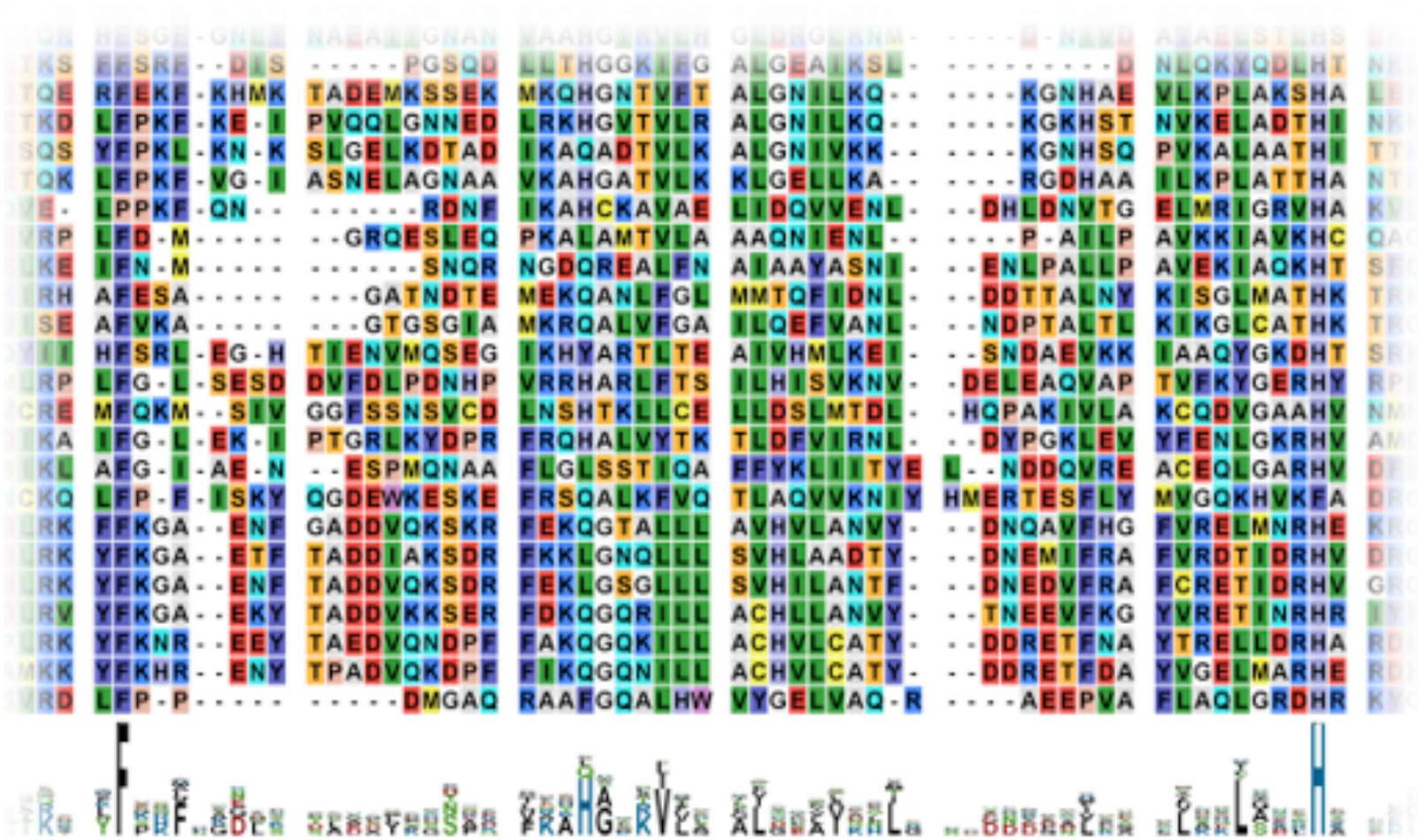
model rodziny białek





- zaczynamy od treningu (czyli estymacji parametrów) alg. Bauma-Welcha,
- długość modelu - średnia długość sekwencji w rodzinie,
- inicjalizacja parametrów - rozkłady jednostajne.

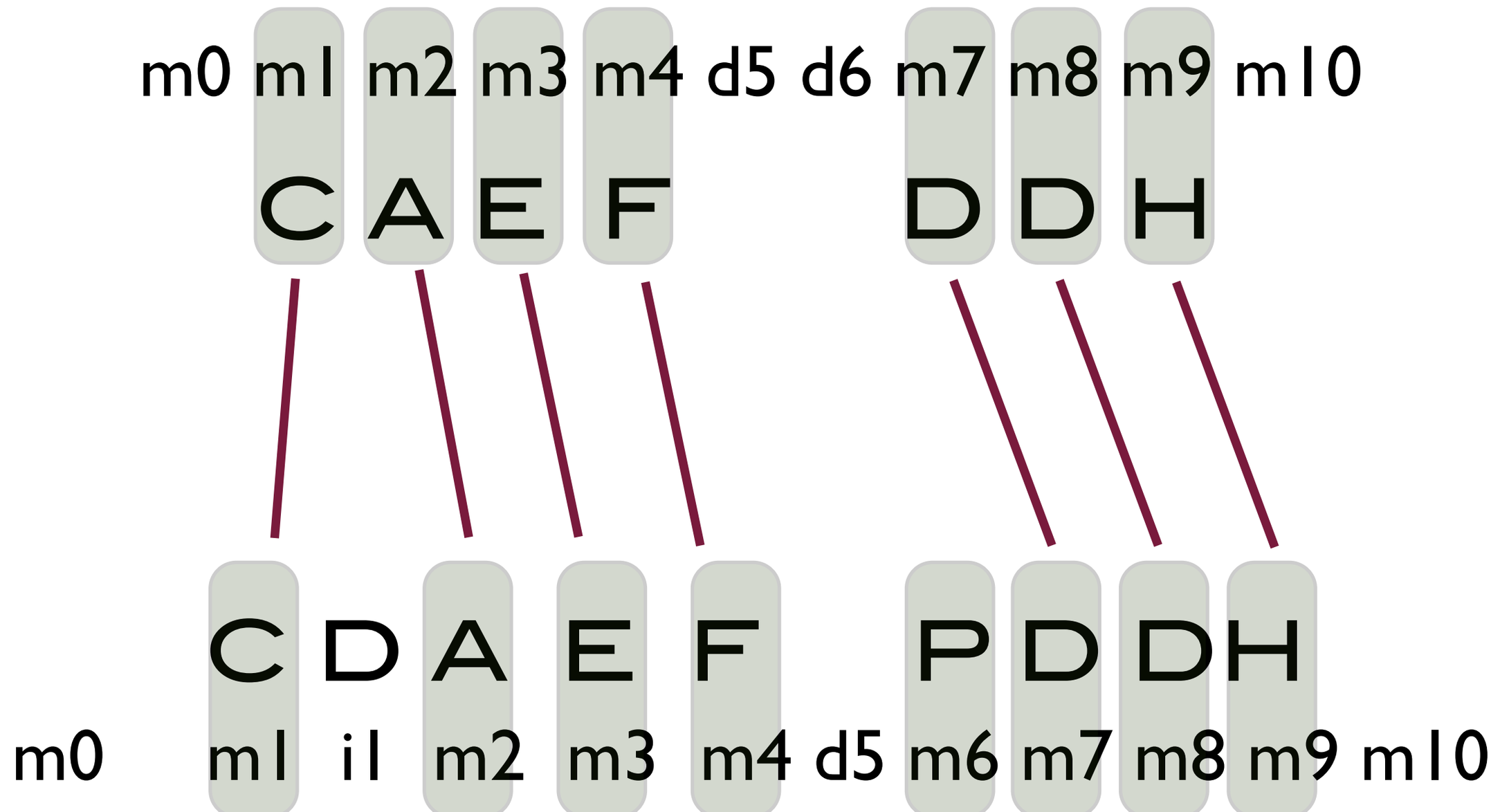
hmm i multiuliniowienia



hmm i multiuliniowienia

- używamy sekwencji do wytrenowania modelu hmm (alg. Bauma-Welcha).
- dla każdej sekwencji algorytmem Viterbiego liczymy najbardziej prawdopodobną trajektorię stanów łańcucha.
- konstruujemy uliniowienie: dwa aminokwasy są w jednej kolumnie jeśli są wyemitowane w tym samym stanie typu match.

hmm i multiuliniowienia



hmm i multiuliniowienia

C - A E F - D D H

C D A E F P D D H

hmm i multiuliniowienia

C A E F T P A V H
C K E T T P A D H
C A E T P D D H
C A E F D D H
C D A E F P D D H

hmm i multiuliniowienia

m0 m1 m2 m3 m4 m5 m6 m7 m8 m9 m10

m0 m1 m2 m3 m4 m5 m6 m7 m8 m9 m10

m0 m1 m2 m3 **d4** m5 m6 m7 m8 m9 m10

m0 m1 m2 m3 m4 **d5 d6** m7 m8 m9 m10

m0 m1 **i1** m2 m3 m4 **d5** m6 m7 m8 m9 m10

hmm i multiuliniowienia

```
C- AEFTPAVH
C- KETTPADH
C- AE-TPDDH
C- AEF--DDH
CDAEF-PDDH
```

hmm i multiuliniowienia

- możliwe niejednoznaczności, np.

dla modelu długości 2 i sekwencji:

A B A C
A B B A C

o trajektoriach:

m0 m1 i1 i1 m2 m3
m0 m1 i1 i1 i1 m2 m3

A B A C
A B B A C

- w takich przypadkach fragmenty uliniowienia małymi literkami kodowane ...

muliuliniowienie globin

- 400 białek globinowych zostało użytych do wytrenowania modelu.
- używając algorytmu Viterbiego uliniowiona całą rodzinę liczącą 625 białek globinowych.
- uzyskane uliniowienie porównane z uliniowieniem strukturalnym (które było znane dla 7 sekwencji).

Pfam

- http://pfam.sanger.ac.uk/
- potrafi namierzyć domeny funkcyjne w sekwencjach białek - reprezentowane jako HMMy (ponad 2000)
- jak annotować nową sekwencję ?
  - BLASTem znajdujemy homologi
  - czasem sensowniej przeszukać domeny z PFAMa

polowanie na geny

- miliardy par zasad DNA do analizy: sekwencje genów poprzedzielane długimi „niefunkcjonalnymi" fragmentami.
- metody obliczeniowe automatycznej detekcji regionów kodujących bardzo pożądane.
- **Genscan**: popularne narzędzie oparte o HMMy
- Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268,** 78-94.

semihidden Markov models

p prawdopodobieństwo pozostania w stanie

$p^{n-1}(1-p)$ p-stwo, że czekamy n kroków

rozkład geometryczny

semiHMM:

1. bez pętli

2. w stanie emitujemy całe słowo z dowolnego rozkładu

semihidden Markov models

z każdym stanem S związana zmienna losowa

L_S - rozkład długości słowa

$Y_{S,l}$ - rozkład na sekwencjach długości l

- algorytmy dla semiHMM dużo bardziej skomplikowane, bo musimy nie tylko odkryć

trajektorię stanów które wyemitowały daną sekwencję, ale również punkty podziału...

semihidden Markov models

- uogólniony algorytm Viterbiego nie działa w rozsądnym czasie... dodatkowe założenia niezbędne:
  - długości długich regionów międzygenowych mają rozkład geometryczny.
  - sekwencje w tych regionach typu iid.

semihidden Markov models

- DEF: parsowanie ϕ to sekwencja stanów $q_1, g_2, \ldots q_r$ i długości $d_1, d_2, \ldots d_r$

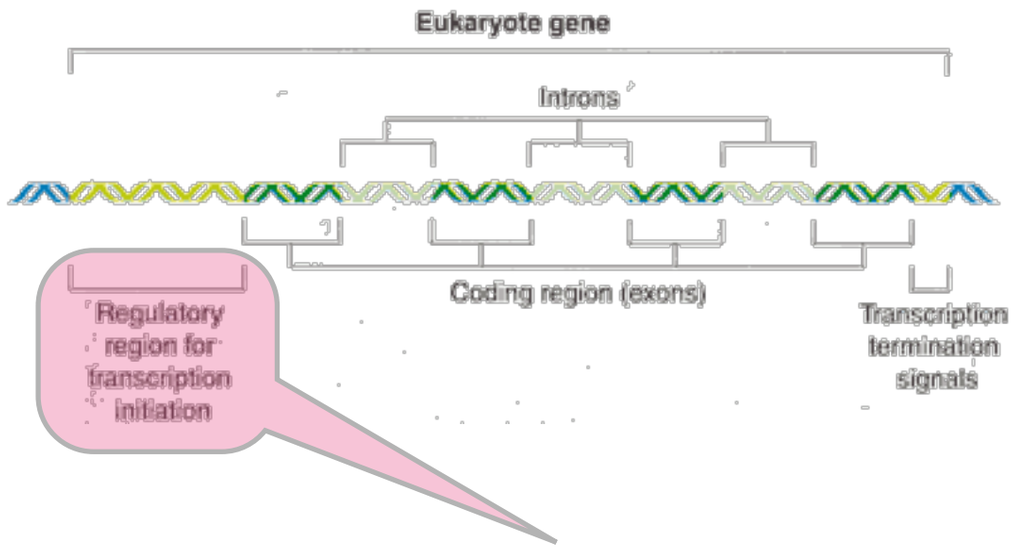
- dla danej sekwencji s algorytm Viterbiego znajduje optymalne parsowanie ϕ_{opt} takie, że

$$P(\phi_{\text{opt}}|s) \geq P(\phi|s)$$

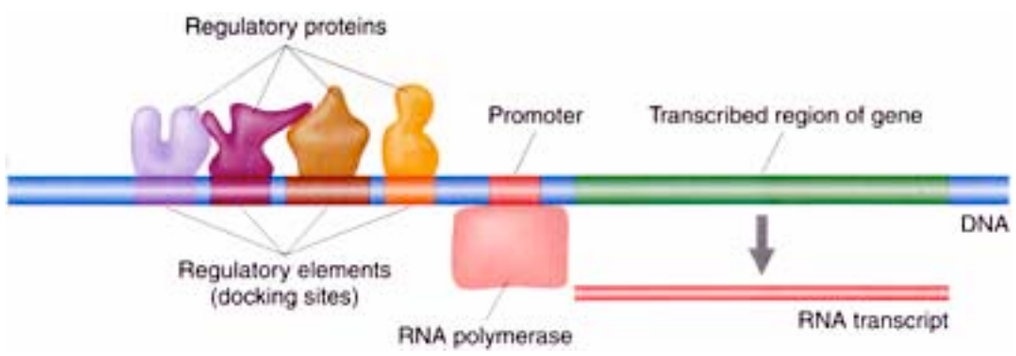
dla wszystkich parsowań ϕ

czyli ϕ_{opt} z największym prawdopodobieństwem wyemitowało sekwencję s

struktura genu - powtórka

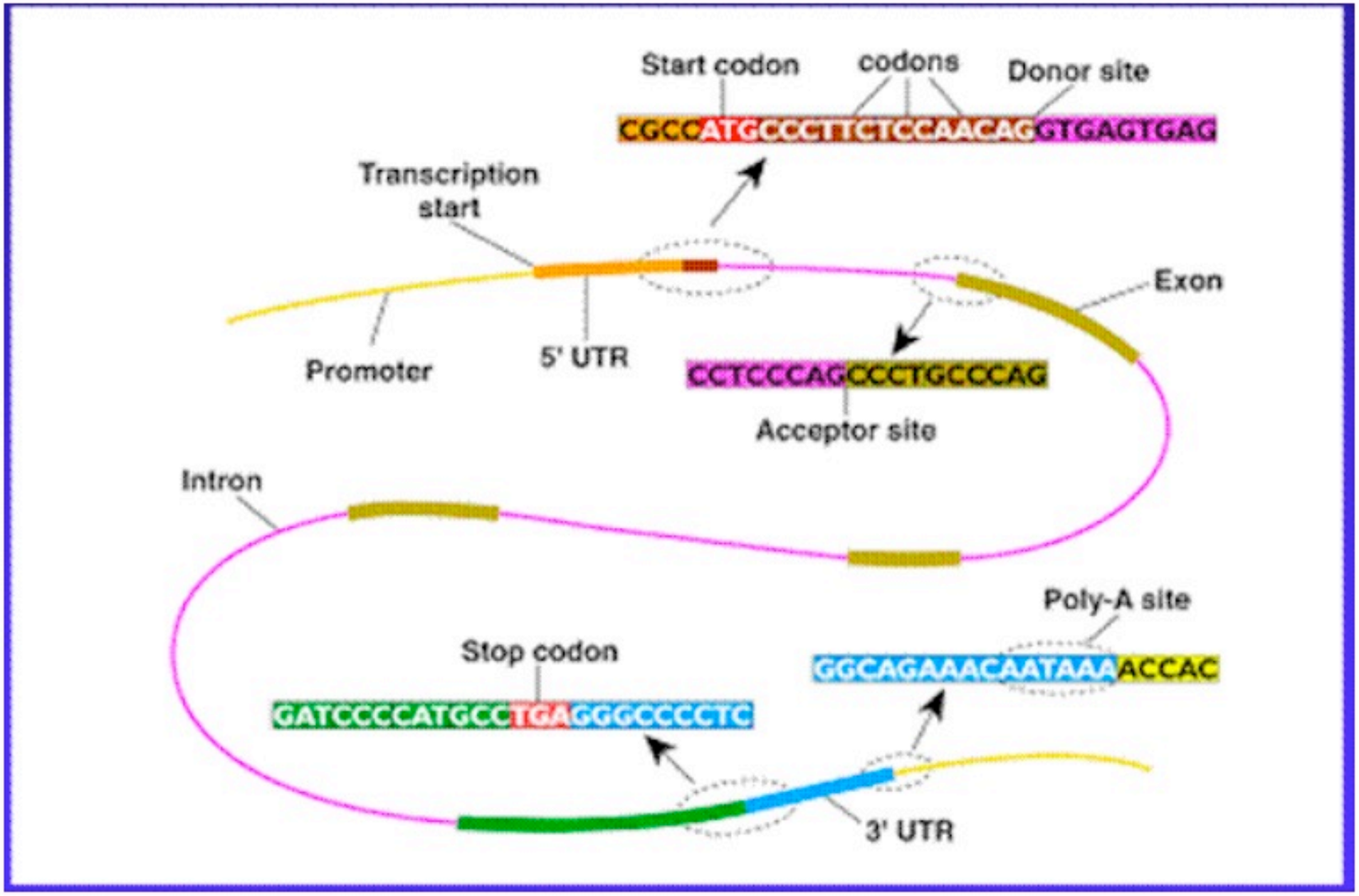


długość około 500 bp



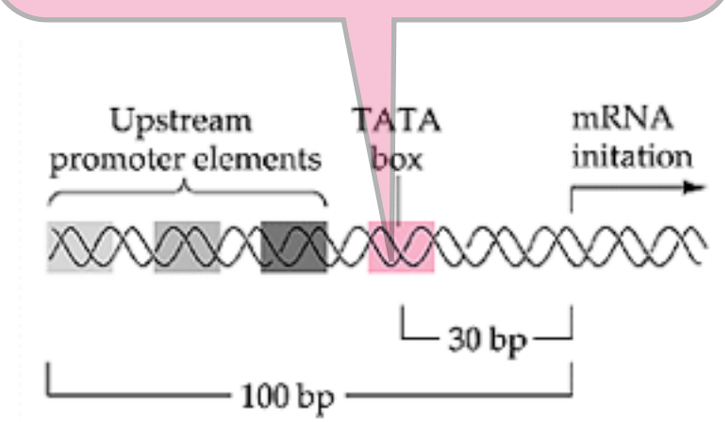
region UTR podlega transkrypcji, ale nie translacji

struktura genu - powtórka

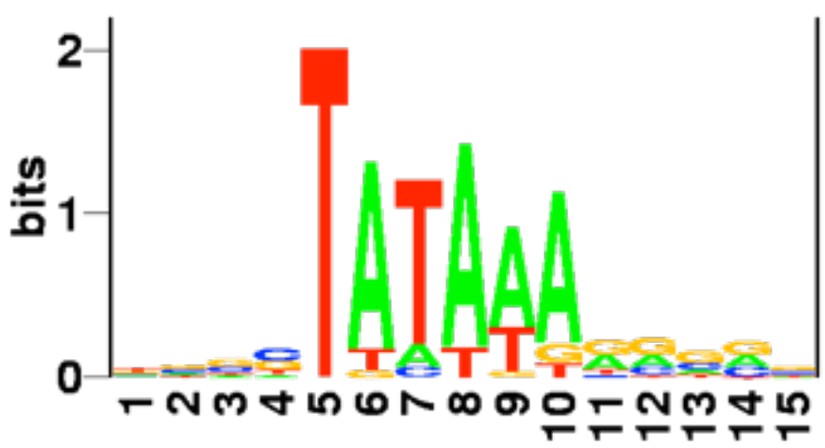


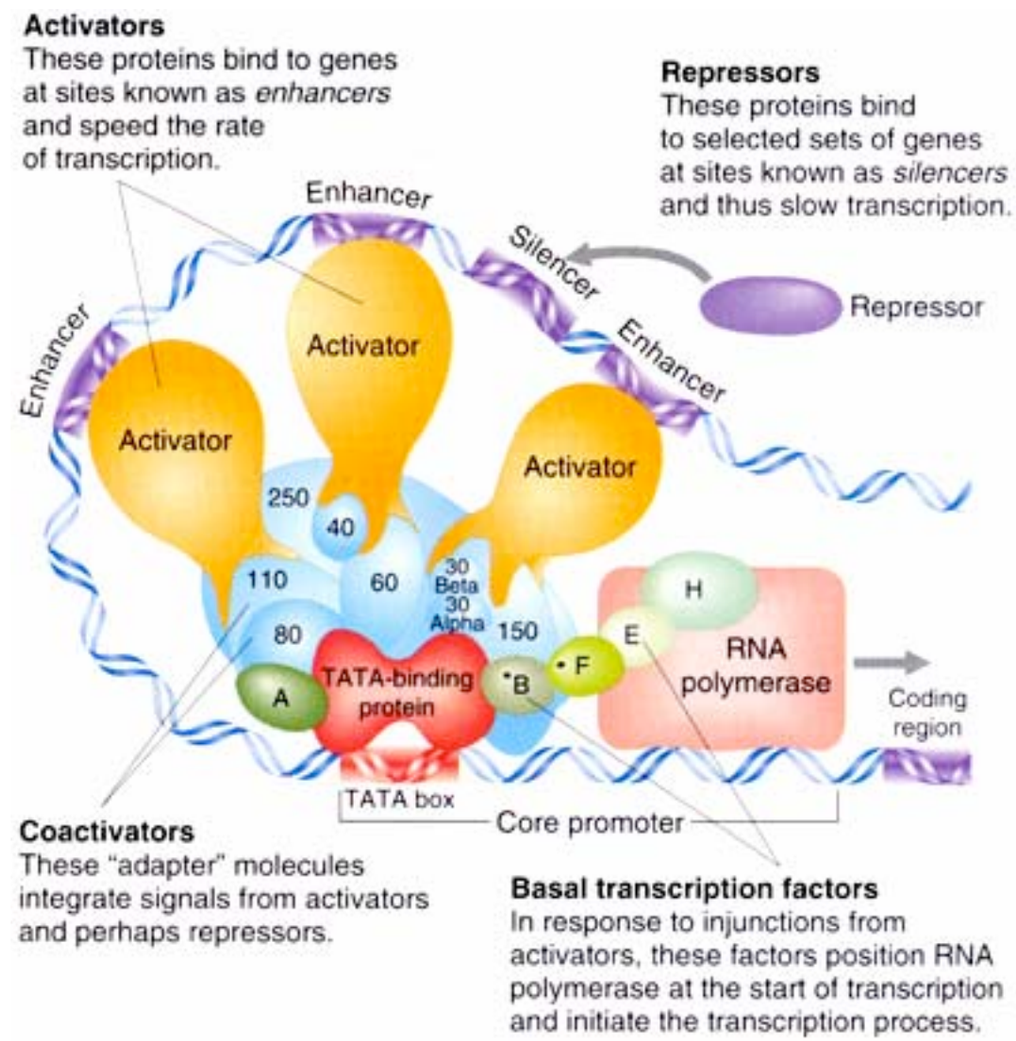
struktura genu - powtórka

ma go 70% genów

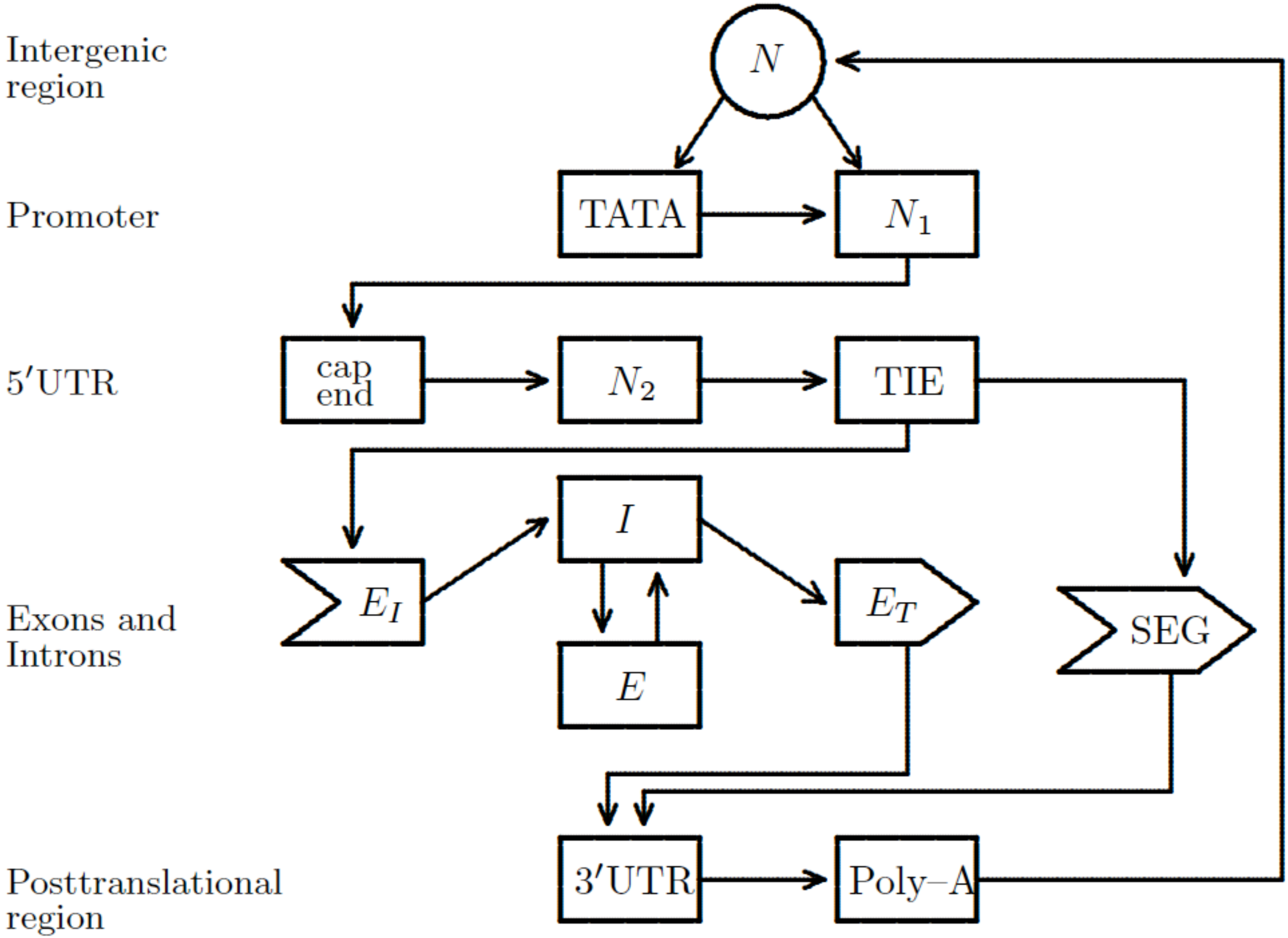


CCAAT
GGGGCGG
GCCACACCC
ATGCAAAT





Scientific American image

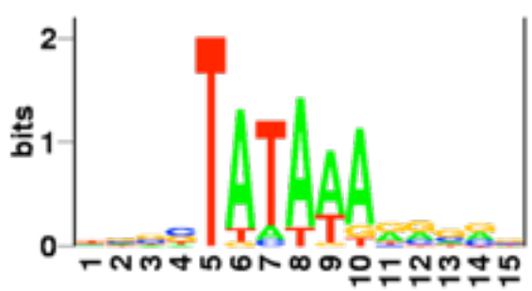

Intergenic region
Promoter
5′UTR
Exons and Introns
Posttranslational region
N
TATA
N1
cap end
N2
TIE
I
EI
ET
E
SEG
3′UTR
Poly–A

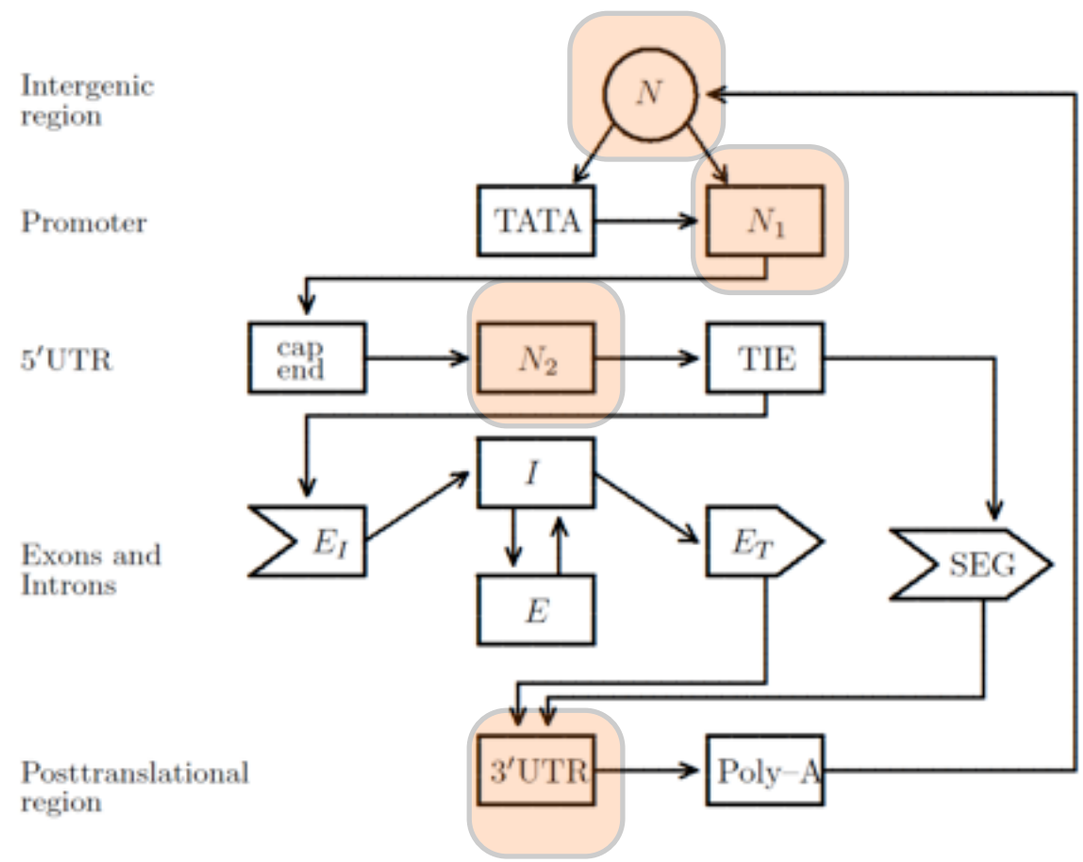
L_N r. geometryczny

sekwencja dł. ℓ generowana przez ŁM rzędu 5

$$3 \cdot 4^5 = 3072 \text{ parameters}$$

intergenic null model





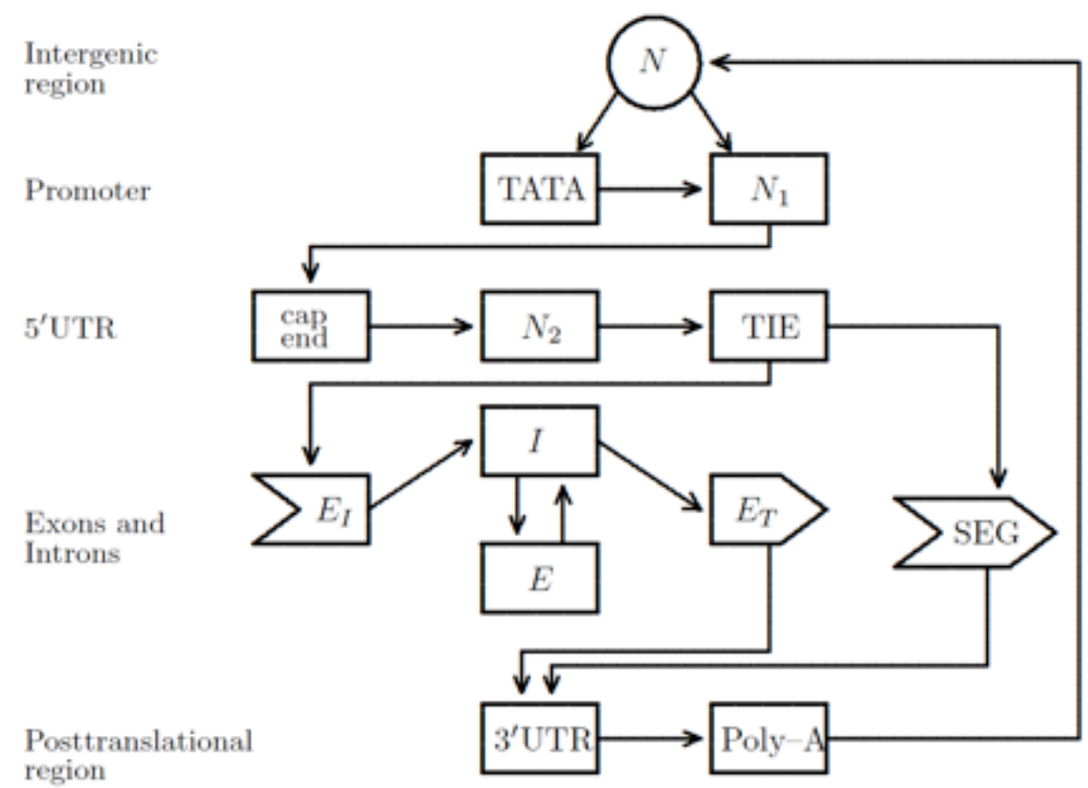
L_{N_1} jednostajny: 28-34

N_2 geometryczny: średnia 735

cap end

TIE

SEG
single exon genes



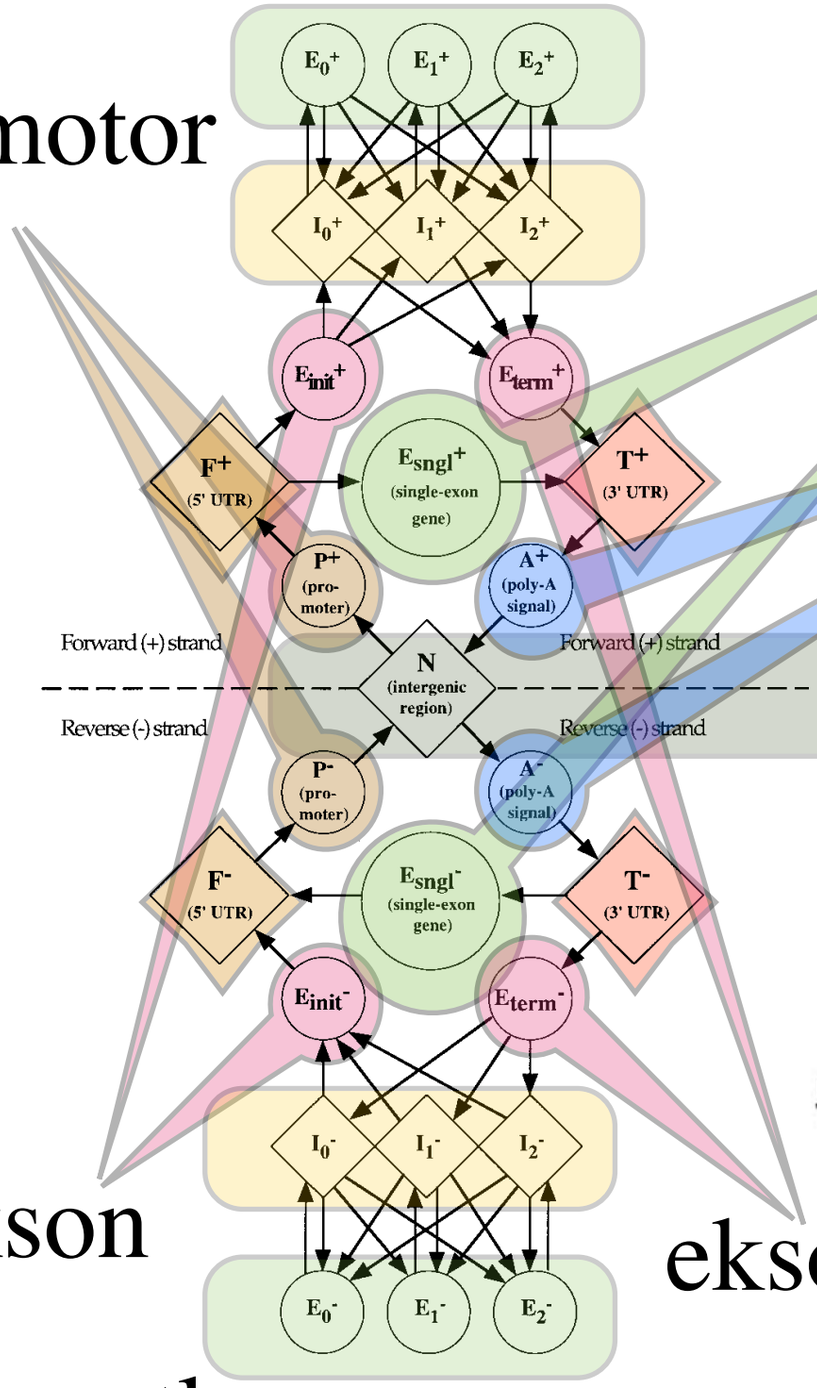
genscan

promotor

pojedynczy ekson

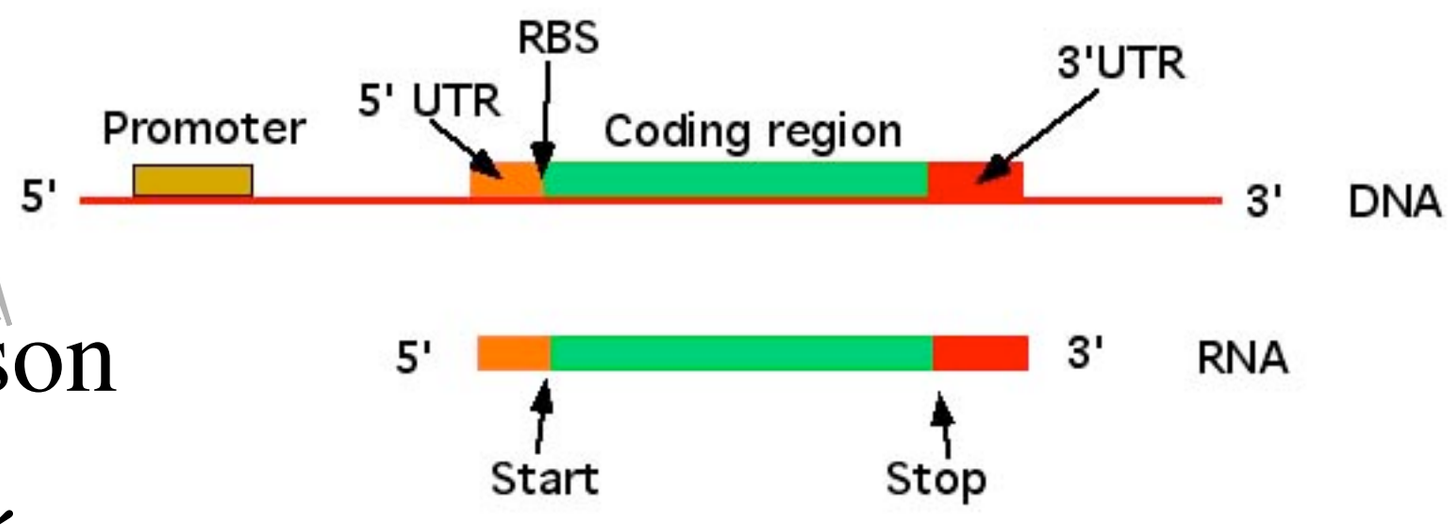
sygnał polyA = koniec transkrypcji

region międzygenowy



ekson początkowy

ekson końcowy



introny/eksony zaczynają się po 1,2,3 bazie w kodonie ....

dla nici komplementarnej

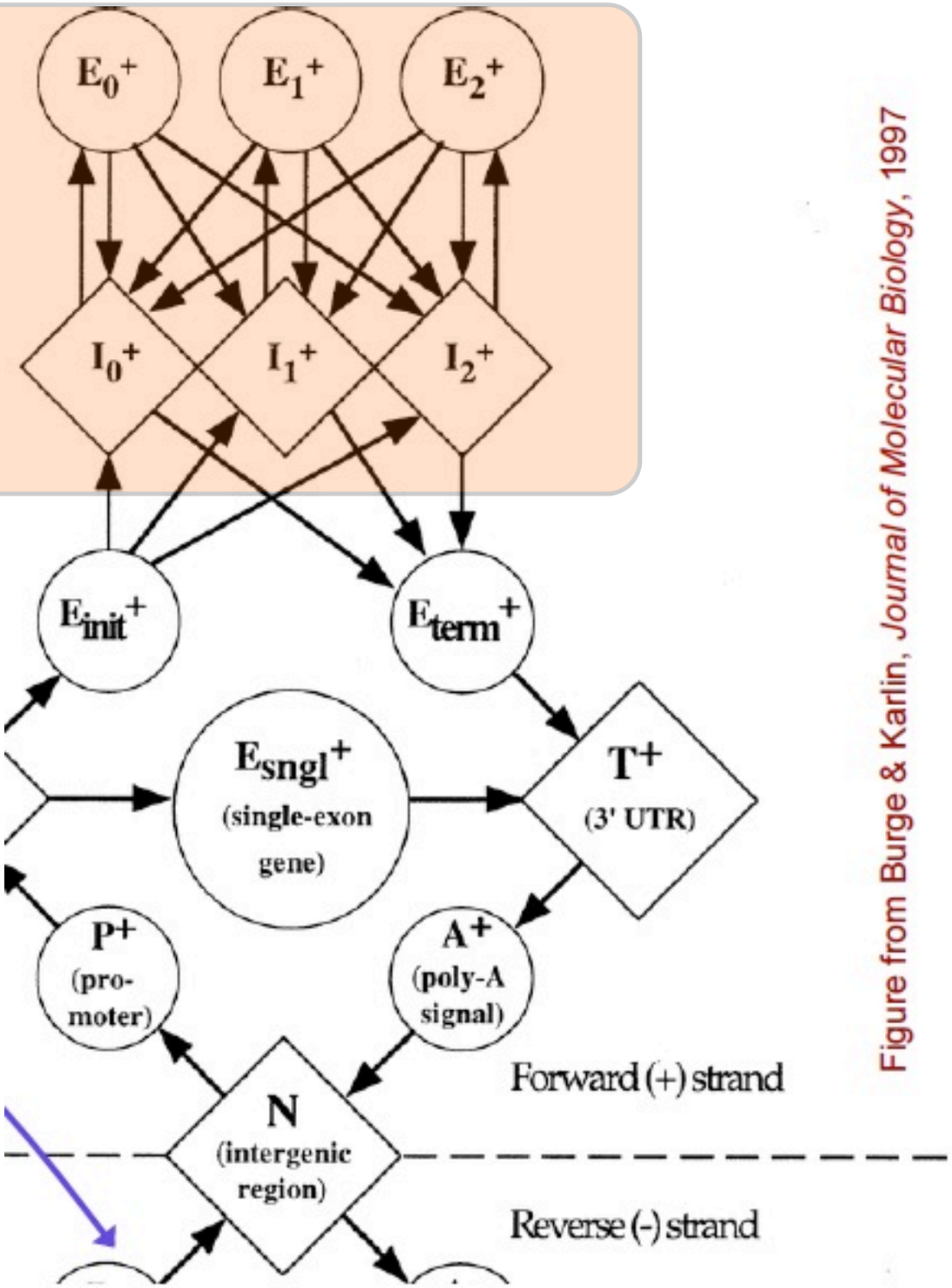
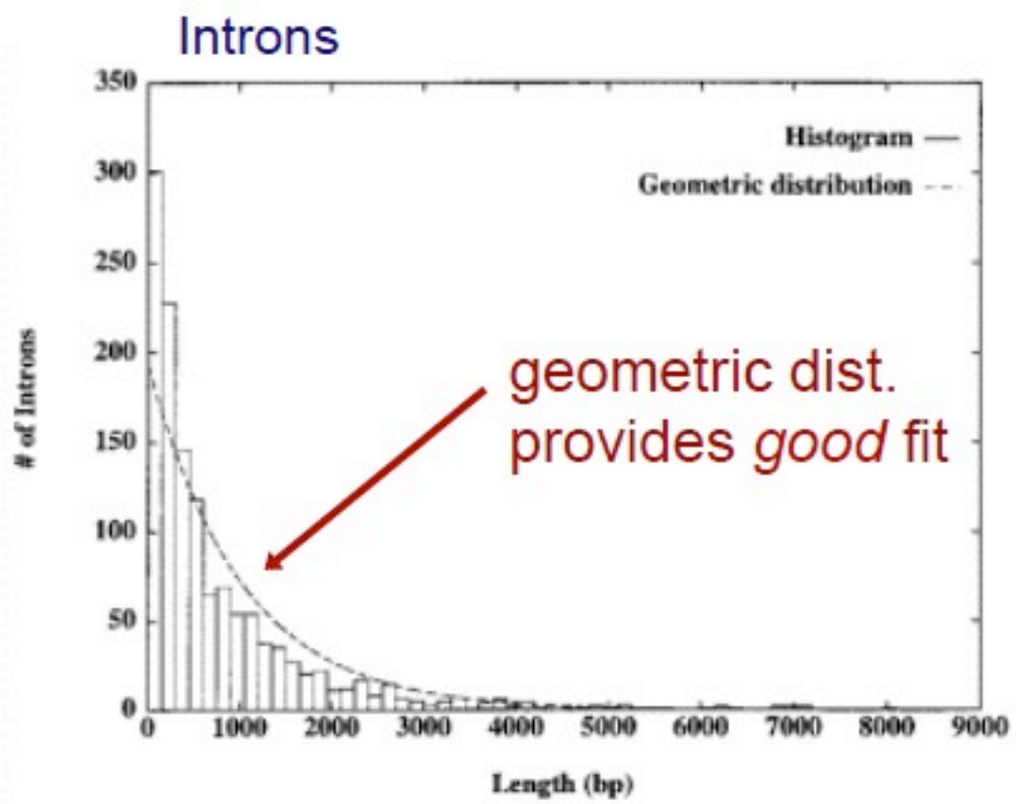
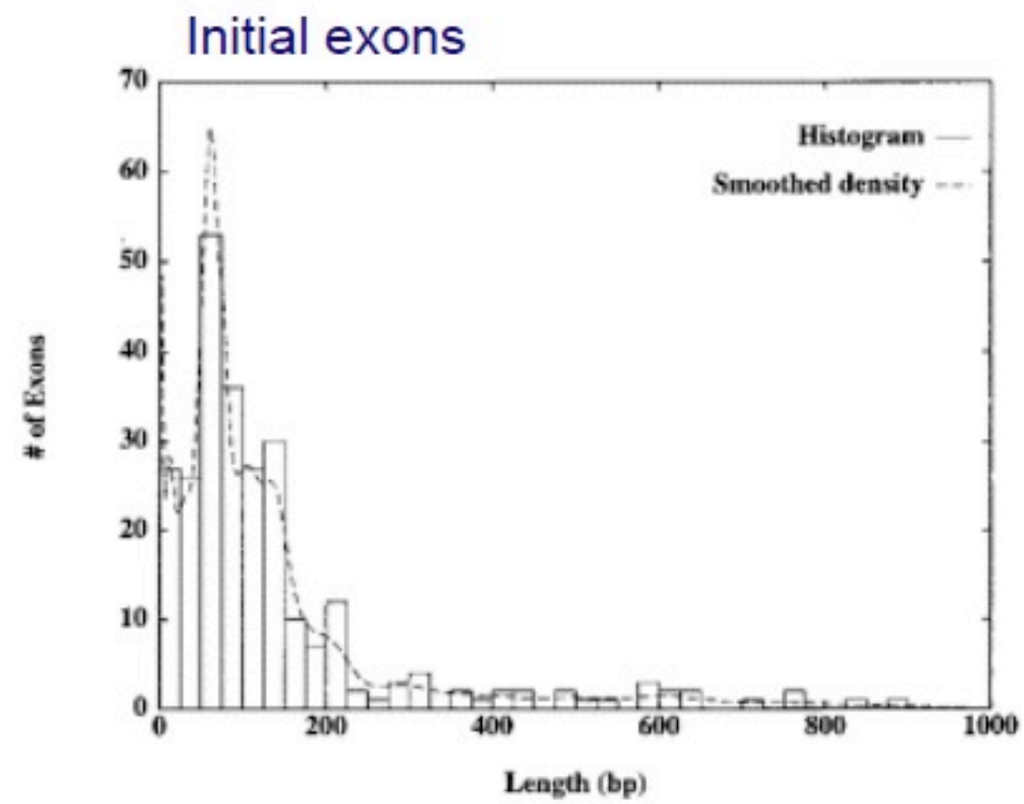
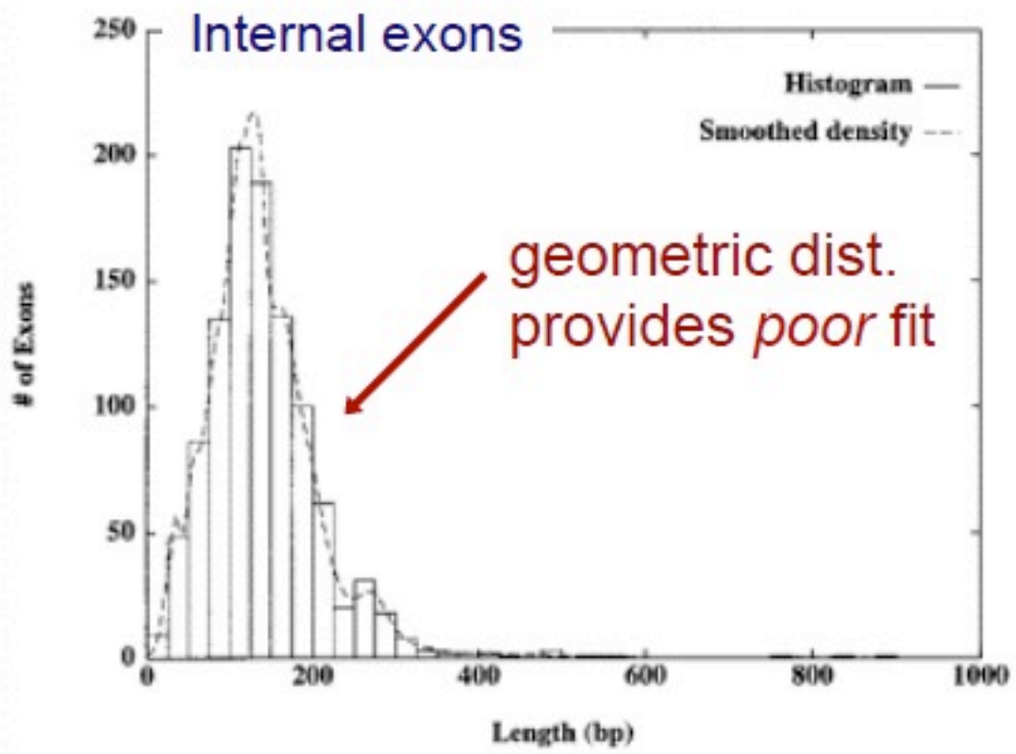


Figure from Burge & Karlin, *Journal of Molecular Biology*, 1997

- dla każdego typu sekwencji GENSCAN modeluje
  - rozkład długości: nieparametrycznie - histogram lub parametrycznie - r. geometryczny;
  - skład sekwencji: łańcuch Markowa (1st, 5 stopnia, (nie)homogeniczny







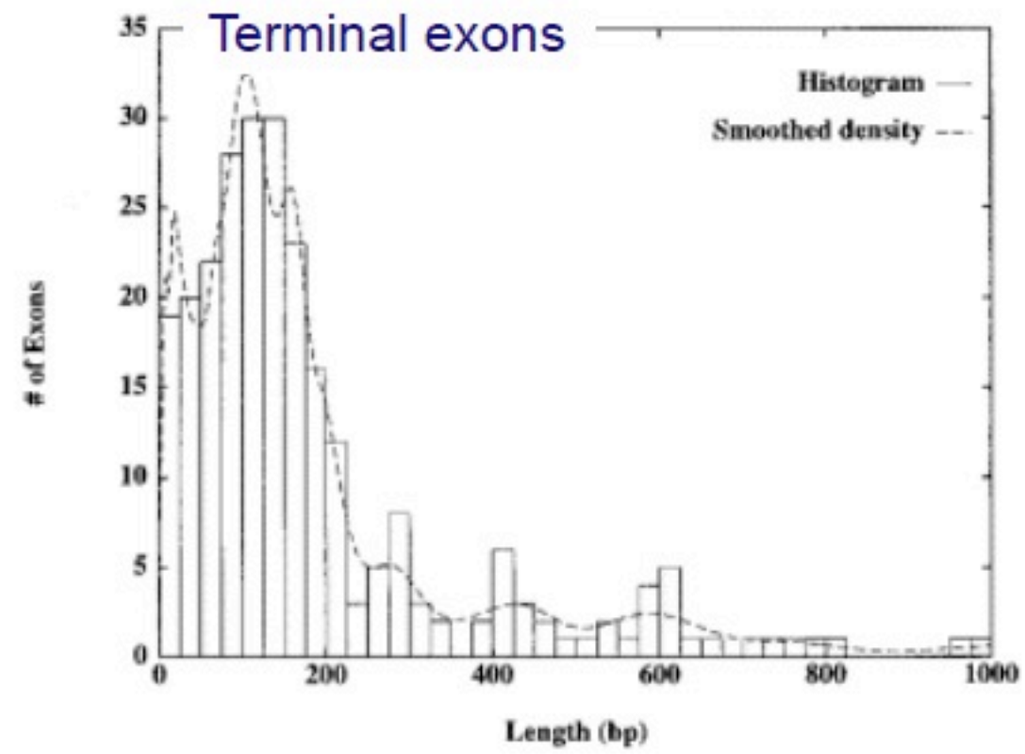
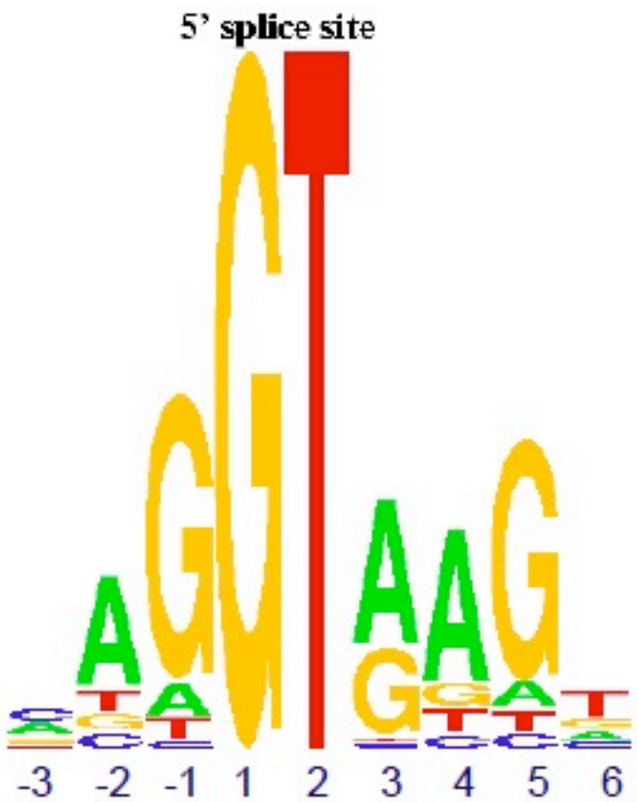


Figure from Burge & Karlin, *Journal of Molecular Biology*, 1997

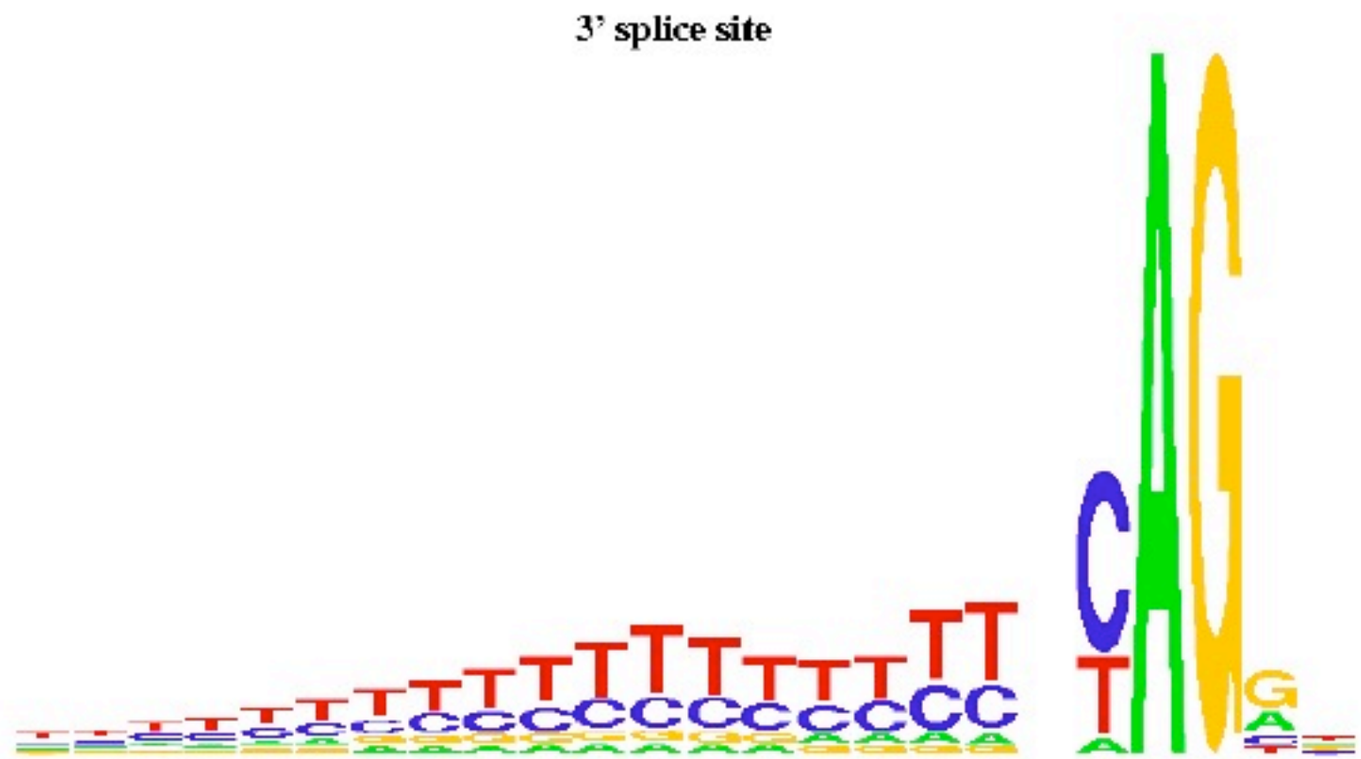
Splice Signals

donor sites



exon

acceptor sites

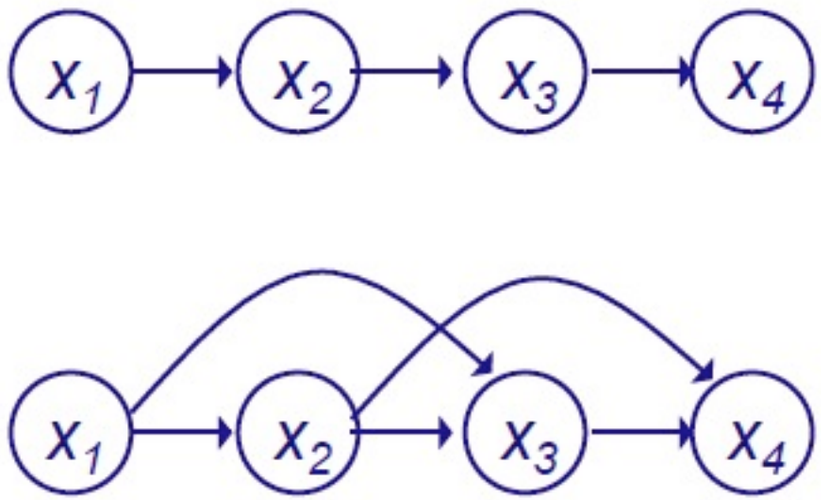


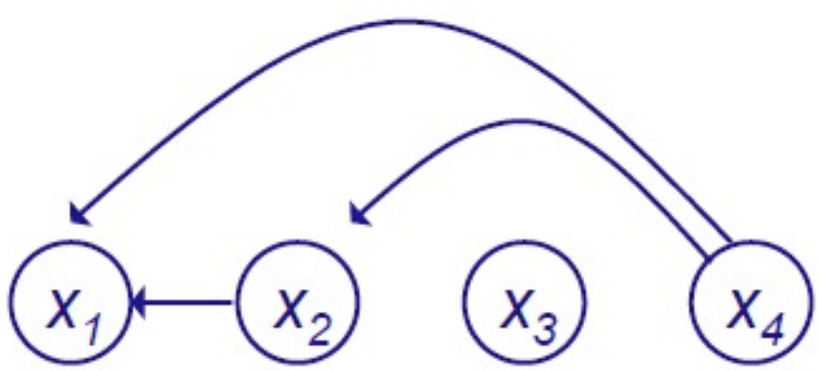
exon

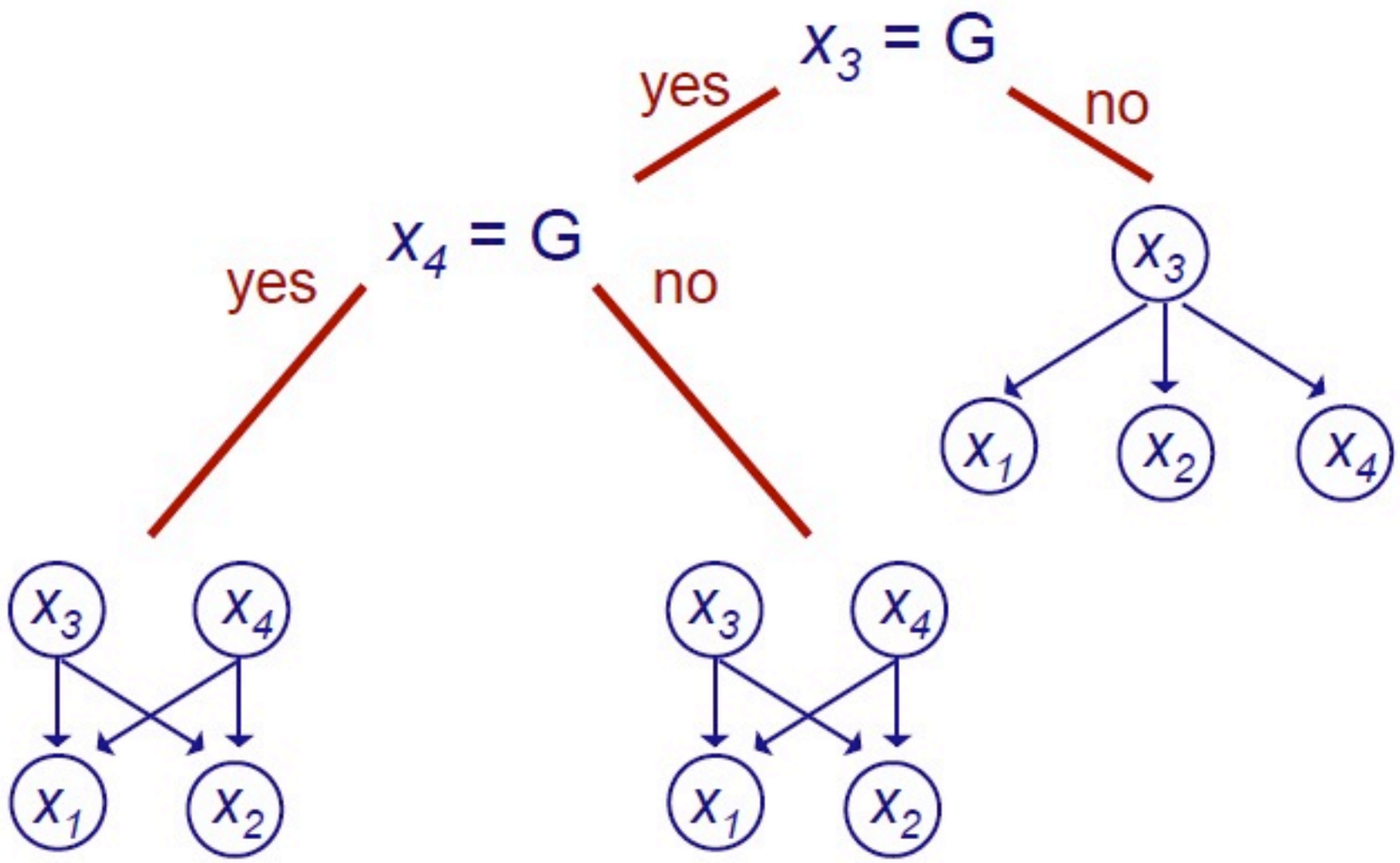
Figures from Yi Xing

zależności między kolumnami...

jak modelować ?






x_3 = G
yes
no
x_4 = G
yes
no
x_3
x_1
x_2
x_4
x_3
x_4
x_1
x_2
x_3
x_4
x_1
x_2

Test asocjacyjny chi kwadrat

Często dane, które badamy mają postać tabeli, której każda komórka jest zmienną losową:

	1	2	3	$\dots$	c	$\sum$
1	Y_{11}	Y_{12}	Y_{13}	$\dots$	Y_{1c}	$y_{1.}$
2	Y_{21}	Y_{22}	Y_{23}	$\dots$	Y_{2c}	$y_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	Y_{r1}	Y_{r2}	Y_{r3}	$\dots$	Y_{rc}	$y_{r.}$
$\sum$	$y_{.1}$	$y_{.2}$	$y_{.3}$	$\dots$	$y_{.c}$	y

Przy założeniu, że hipoteza zerowa jest poprawna, czyli kategorie wierszy i kolumn są od siebie niezależne, możemy obliczyć oczekiwaną liczbę obserwacji w komórce (j, k), czyli wartość oczekiwaną zmiennej losowej Y_{jk}:

$$E_{jk} = E(Y_{jk}) = \frac{y_{j.}y_{.k}}{y}$$

Jeśli hipoteza zerowa jest prawdziwa, to obserwowane wartości zmiennych Y_{jk} powinny być bliskie oczekiwanym, czyli znowu liczymy statystykę chi-kwadrat:

$$\sum_{jk} \frac{(Y_{jk} - E_{jk})^2}{E_{jk}}$$

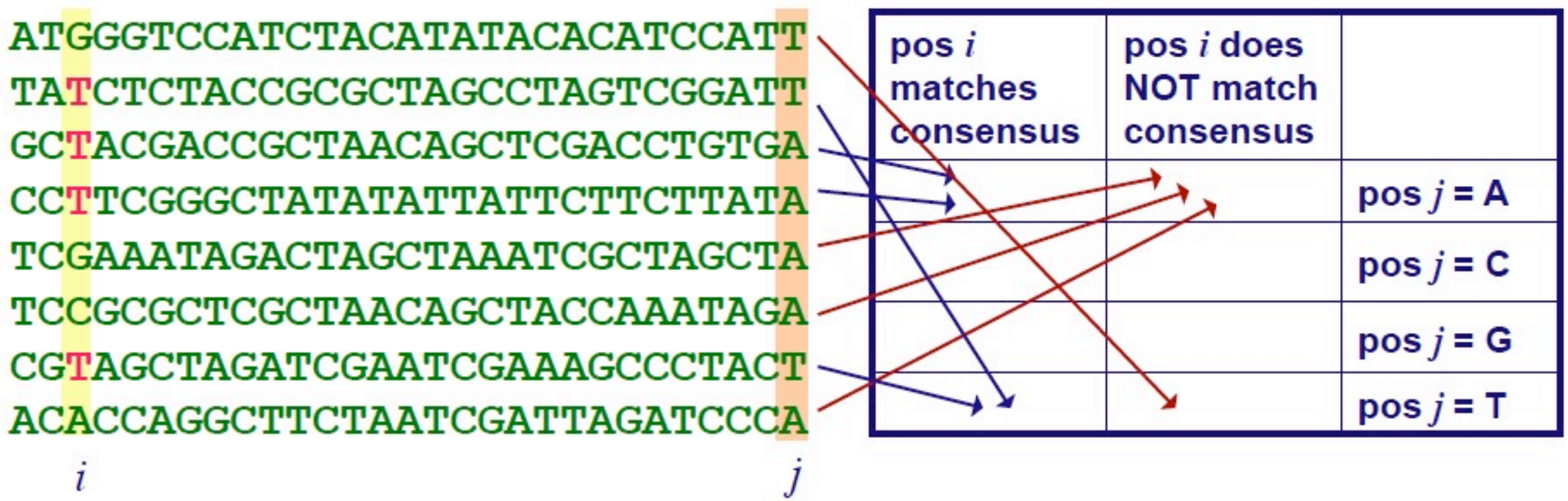
która ma asymptotycznie rozkład chi kwadrat o $\nu = (r-1)(c-1)$ stopniach swobody.

Przykład 3 *Jako przykład rozważmy test statystyczny na niezależność Markowa, który sprowadzi się do testu asocjacyjnego w tabeli* $4 \ x \ 4$ *i będzie badał, czy częstość występowania danego nukleotydu na danej pozycji zależy od rodzaju nukleotydu na pozycji sąsiedniej.*

Przeprowadza się taki test, żeby ocenić czy można modelować sekwencję DNA jako ciąg prób Bernoulliego, czy raczej z użyciem łańcucha Markowa, który uwzględni taką zależność. Tabela asocjacyjna wygląda w naszym przypadku następująco:

	a	c	g	t	$\sum$
a	Y_{11}	Y_{12}	Y_{13}	Y_{14}	$y_{1.}$
c	Y_{21}	Y_{22}	Y_{23}	Y_{24}	$y_{2.}$
g	Y_{31}	Y_{32}	Y_{33}	Y_{34}	$y_{3.}$
t	Y_{41}	Y_{42}	Y_{43}	Y_{44}	$y_{4.}$
$\sum$	$y_{.1}$	$y_{.2}$	$y_{.3}$	$y_{.4}$	y

max. dependence decomposition - MDD



compute χ^2 values using 2×4 table

alternative hypothesis: distribution for column j depends on whether the consensus base is in column i

null hypothesis: distribution for column j is the same in both cases

pozycja 4 ma największy wpływ na inne...

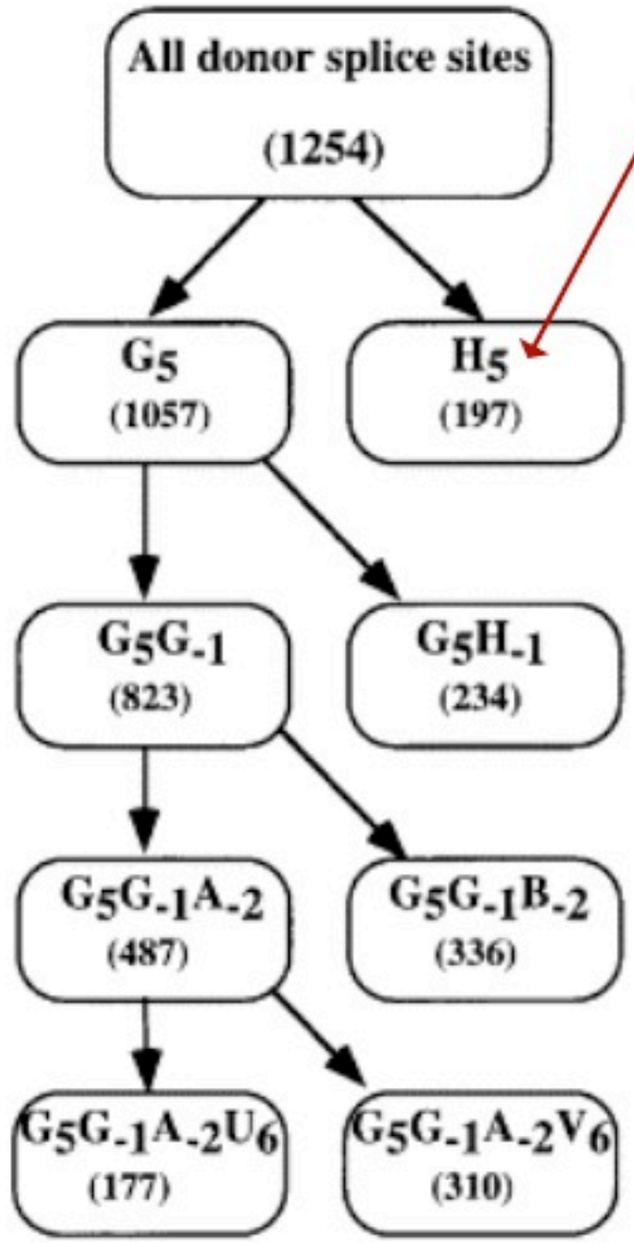
	1	2	3	4	5	total
1		34.2*	7.1	37.2*	2.8	81.3
2	34.2*		.4	72.4*	4.5	111.5
3	7.1	.4		15.3	98.3*	121.1
4	37.2*	72.4*	15.3		14.2	139.1
5	2.8	4.5	98.3*	14.2		119.8

Table 4. Dependence between positions in human donor splice sites: χ^2 statistic for consensus indicator variable C_i *versus* nucleotide indicator X_j

i	Con	j: −3	−2	−1	+3	+4	+5	+6	Sum
−3	c/a	—	61.8*	14.9	5.8	20.2*	11.2	18.0*	131.8*
−2	A	115.6*	—	40.5*	20.3*	57.5*	59.7*	42.9*	336.5*
−1	G	15.4	82.8*	—	13.0	61.5*	41.4*	96.6*	310.8*
+3	a/g	8.6	17.5*	13.1	—	19.3*	1.8	0.1	60.5*
+4	A	21.8*	56.0*	62.1*	64.1*	—	56.8*	0.2	260.9*
+5	G	11.6	60.1*	41.9*	93.6*	146.6*	—	33.6*	387.3*
+6	t	22.2*	40.7*	103.8*	26.5*	17.8*	32.6*	—	243.6*

Pos	A%	C%	G%	U%
-3	33	36	19	13
-2	56	15	15	15
-1	9	4	78	9
+3	44	3	51	3
+4	75	4	13	9
+6	14	18	19	49

Pos	A%	C%	G%	U%
-3	35	44	16	6
-2	85	4	7	5
-1	2	1	97	0
+3	81	3	15	2
+4	51	28	9	12
+6	22	20	30	28



Pos	A%	C%	G%	U%
-3	34	37	18	11
-2	59	10	15	16
+3	40	4	53	3
+4	70	4	16	10
+6	17	21	21	42

Pos	A%	C%	G%	U%
-3	29	31	21	18
-2	43	30	17	11
+3	56	0	43	0
+4	93	2	3	3
+6	5	10	10	76

Pos	A%	C%	G%	U%
-3	37	42	18	3
+3	39	5	51	5
+4	62	5	22	11
+6	19	20	25	36

Pos	A%	C%	G%	U%
-3	29	30	18	23
+3	42	1	56	1
+4	80	4	8	8
+6	14	21	16	49

Pos	A%	C%	G%	U%
-3	32	40	23	5
+3	27	4	59	10
+4	51	5	25	19

Pos	A%	C%	G%	U%
-3	39	43	15	2
+3	46	6	46	3
+4	69	5	20	7

Figure from Burge & Karlin, *Journal of Molecular Biology*, 1997

budujemy drzewo...

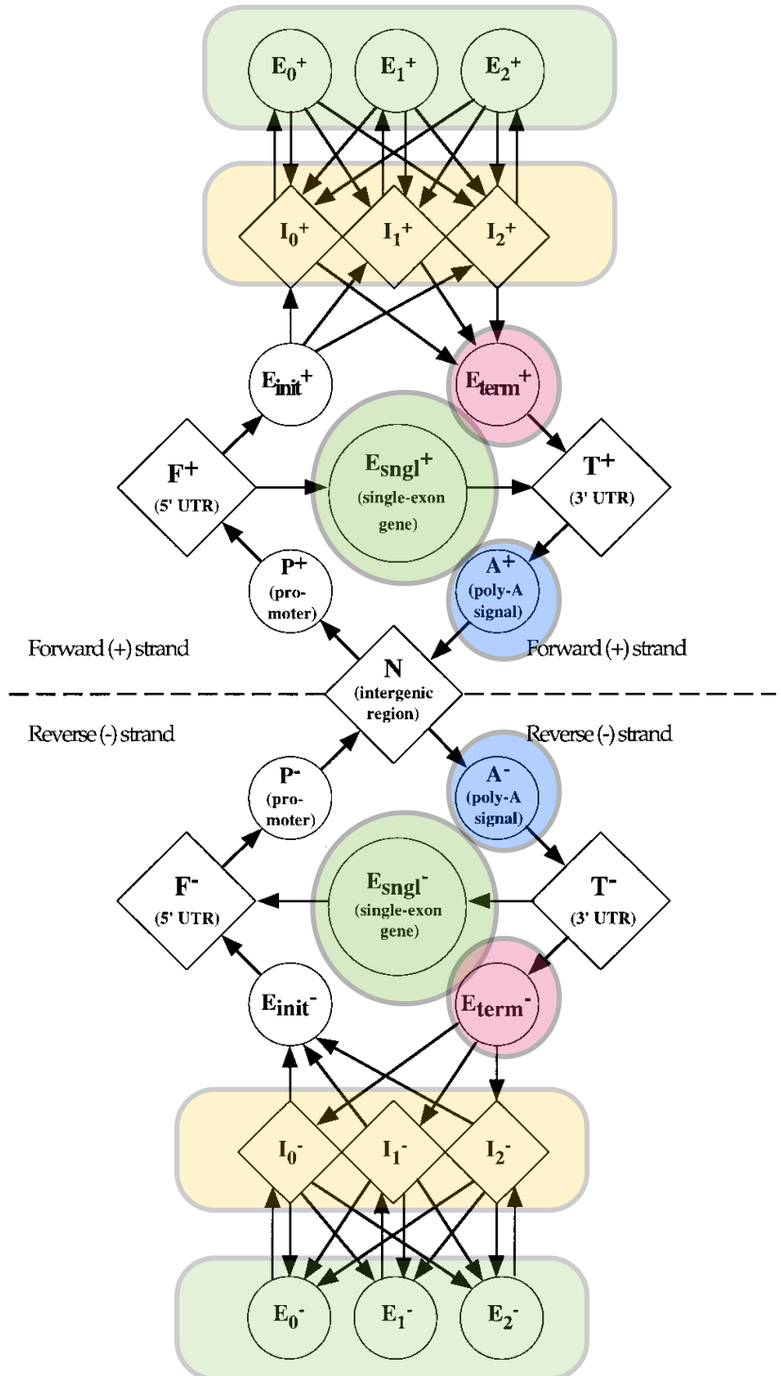
Given: a set of aligned training sequences T
positions $P = \{1, \ldots, k\}$
tree = find_MDD_subtree(T, P)

find_MDD_subtree(T, P)
for each position i *in* P
    determine the consensus base C_i
    calculate dependence between C_i , other positions
if stopping criteria not met
    choose the value of i such that S_i is maximal
    make a node with C_i as the test
    create a single-column PWM for position i
    D_i^+ = sequences in T with base C_i at position i
    D_i^- = other sequences
    left subtree = find_MDD_subtree(D_i^+ , $P - \{ i \}$)
    right subtree = find_MDD_subtree(D_i^- , $P - \{ i \}$)
else
    create a partial PWM for remaining positions in P

test for position j
conditioned on match to
consensus at i

$$S_i = \sum_{j \neq i} \chi^2(C_i, x_j)$$

trening semiHMMa



2.5 Mb bp

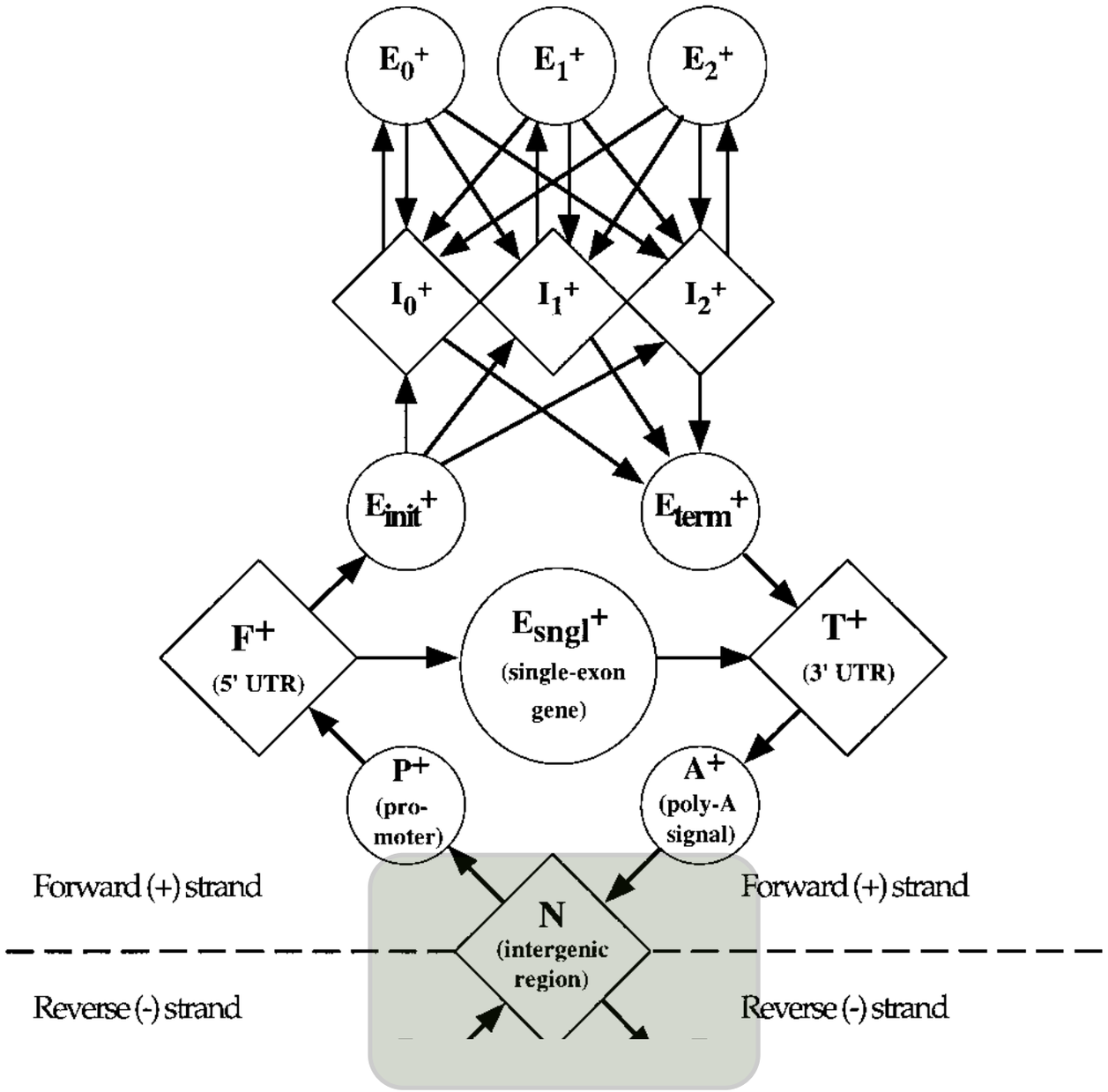
380 genów

142 z pojedynczym eksonem

1492 eksony

1254 introny

genscan



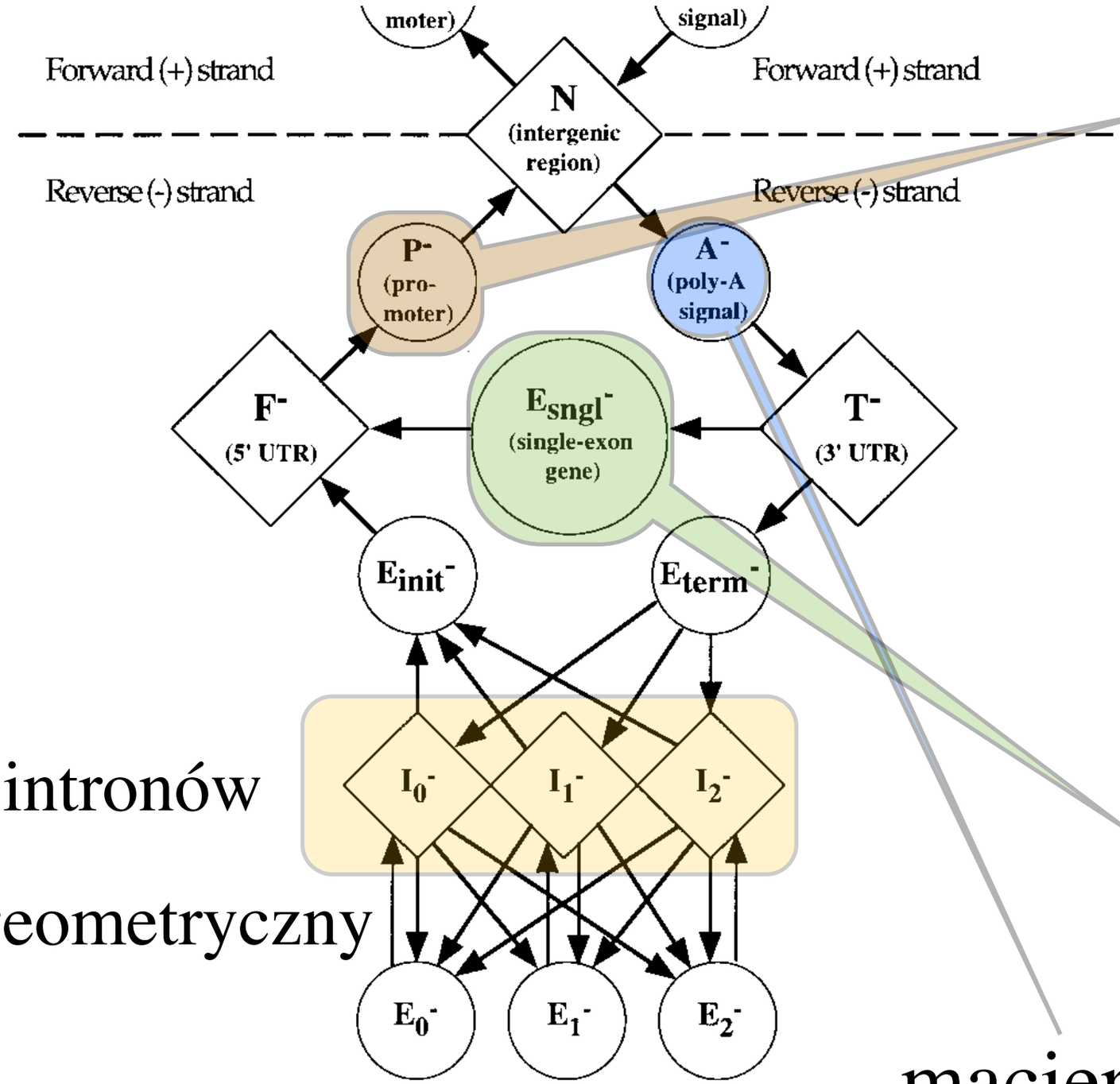
założenie: w danym odcinku liczba genów - rozkład Poissona

odległość między nimi - rozkład wykładniczy

L_N r. geometryczny

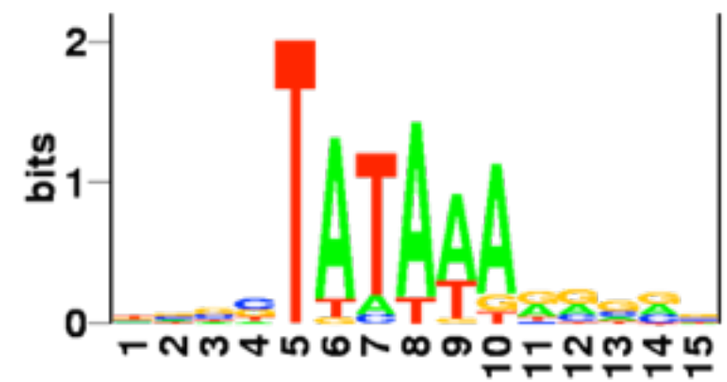
$Y_{N,l}$ generowane przez łańcuch Markowa rzędu 5 (3072 parametry)

genscan



macierz wag

rozmiaru 15



dł. intronów

r. geometryczny

wyuczony łańcuch Markowa rzędu 5

macierz wag

genscan

algorytmem Viterbiego znajdujemy optymalne parsowanie: