Motywy miejsc wiązania czynników transkrypcyjnych

Wykład dla biotechnologów

Bartek Wilczyński

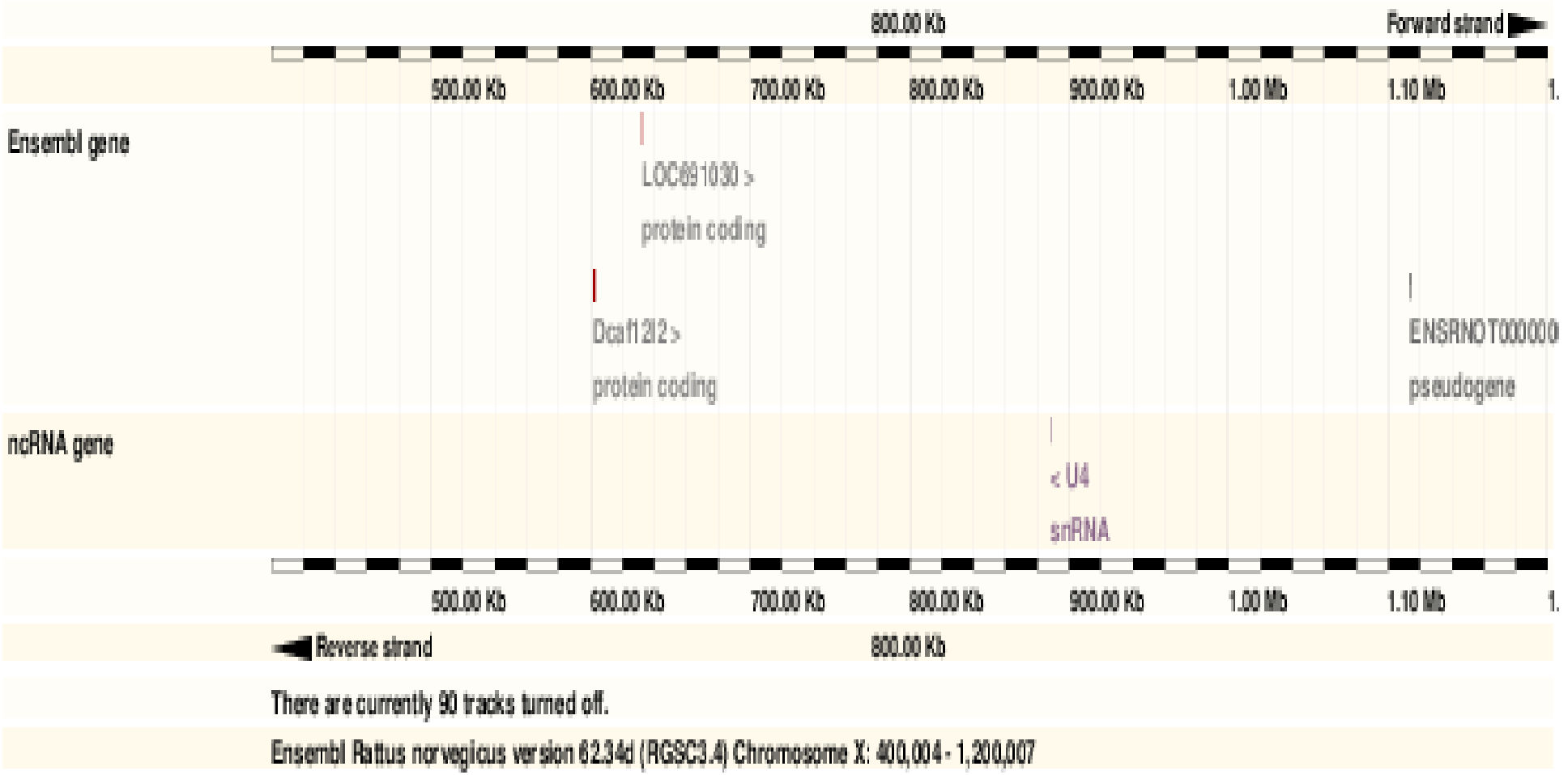
bartek@mimuw.edu.pl

28. listopada 2013

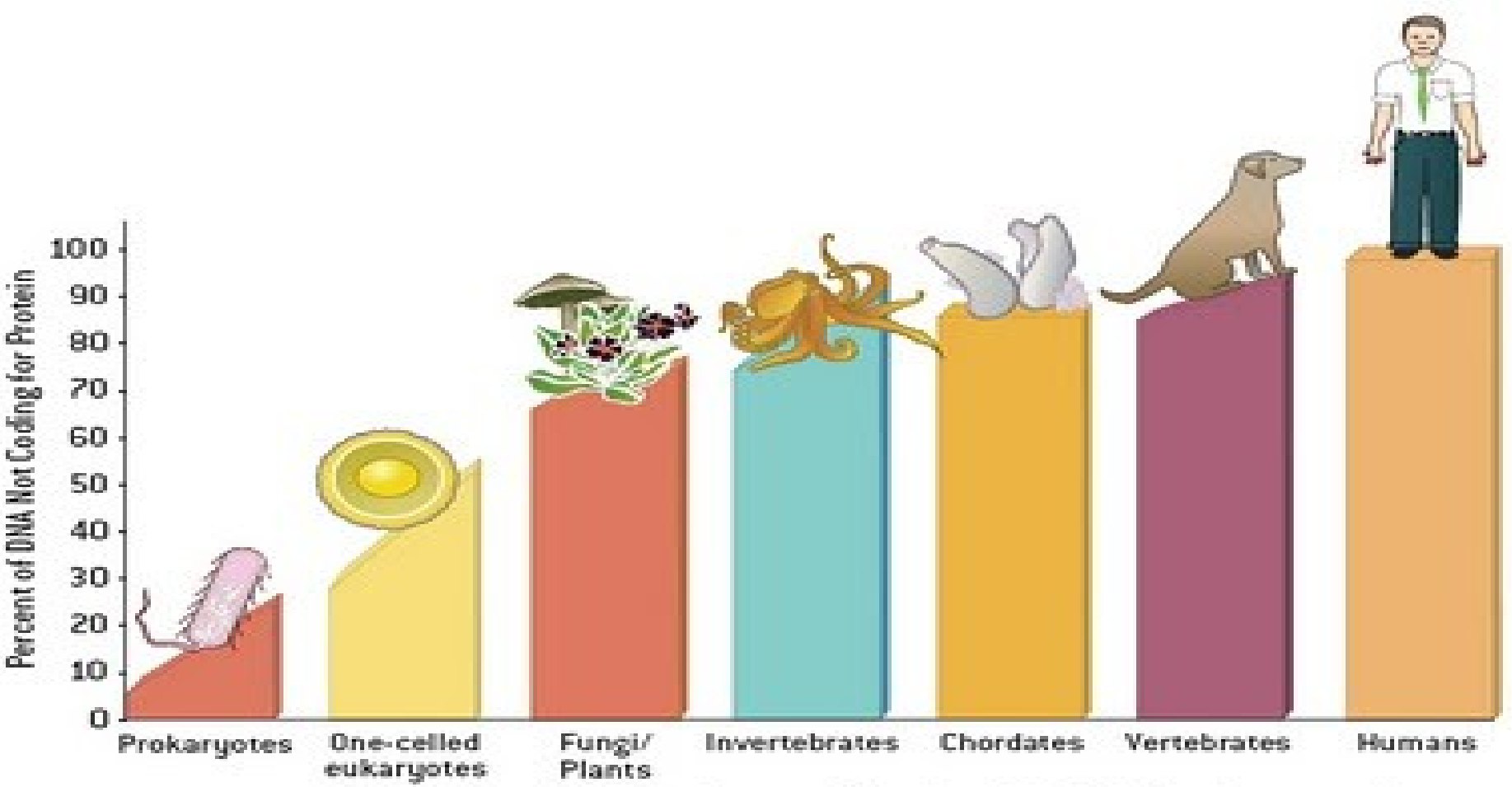
Dlaczego większość genomu nie koduje białek?

- Geny kodujące białka są niewątpliwie ważnymiu elementami genomu
- Dość dobrze rozumiemy rolę genów kodujących białka I reguły rządzące ich ewolucją
- Introny I sekwencje międzygenowe nie mają oczywistej funkcji (zwłaszcza jeśli mowa o fragmentach nie podlegających transkrypcji)
- Analizując profil mutacji w dużych genomach, możemy oszacować, że ok. 30% genomu nie podlega presji selekcyjnej

Losowy fragment genomu ludzkiego...

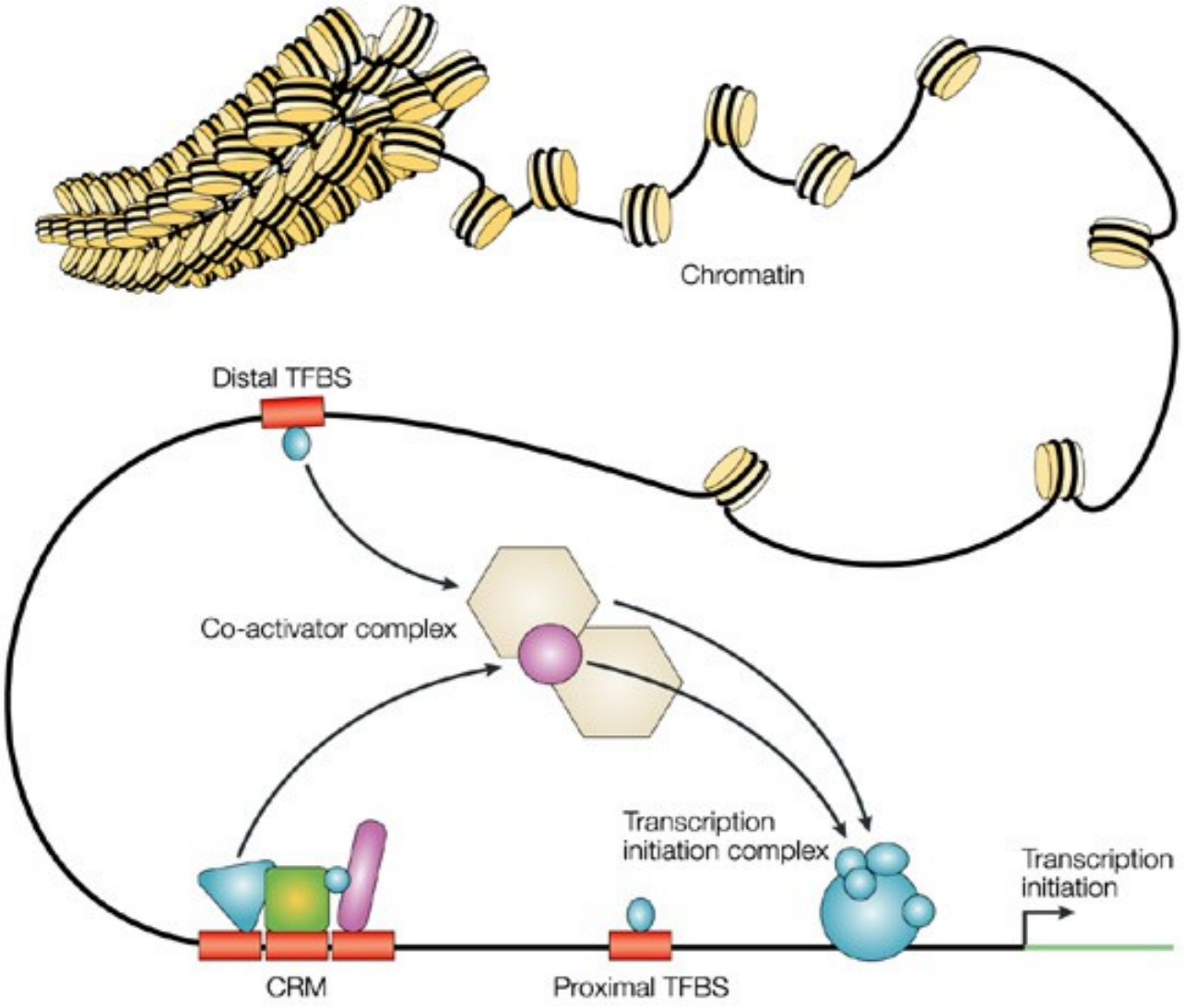


Czy to naprawdę “śmieciowe” DNA?

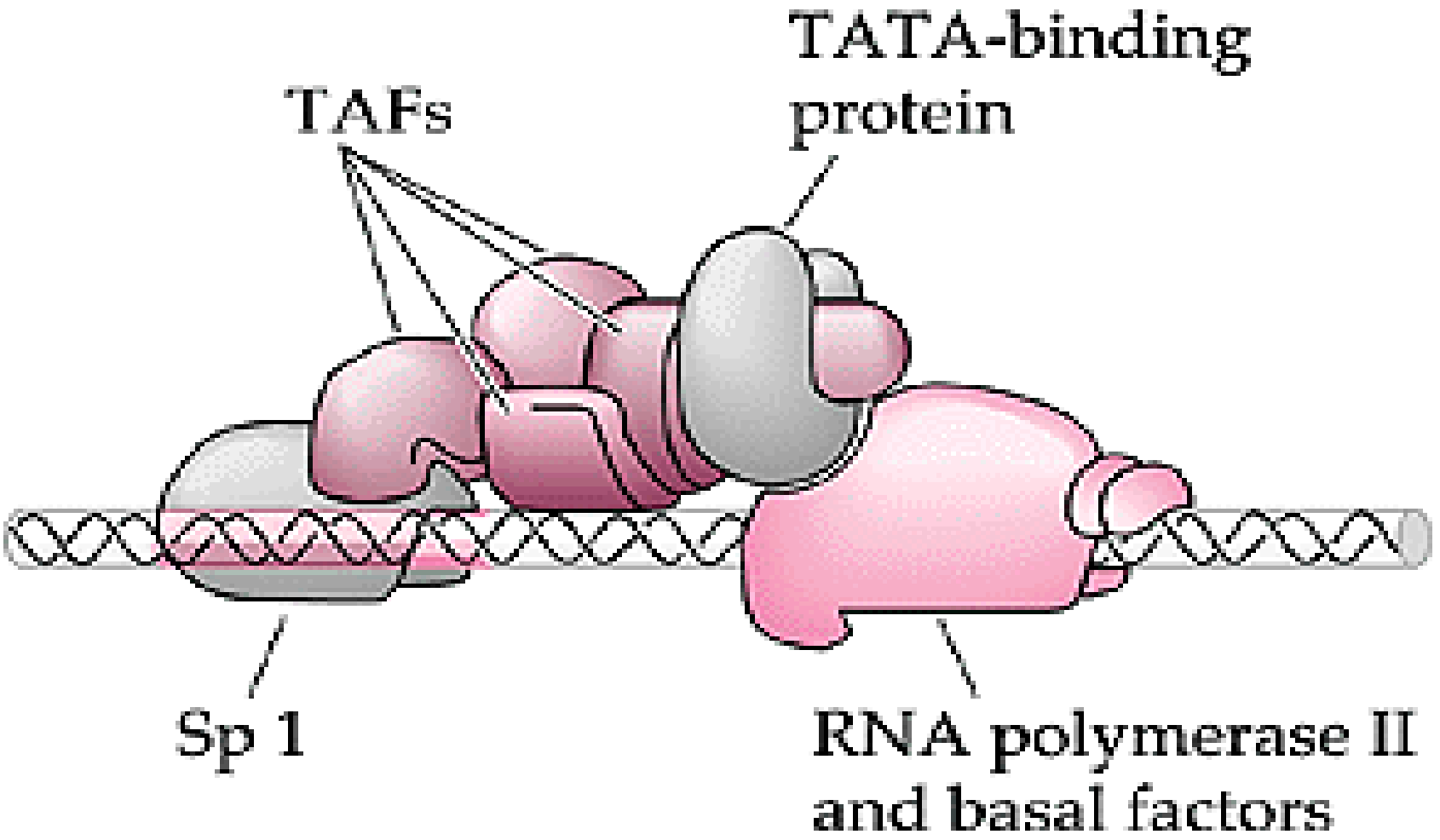


NONPROTEIN-CODING SEQUENCES make up only a small fraction of the DNA of prokaryotes. Among eukaryotes, as their complexity increases, generally so, too, does the proportion of their DNA that does not code for protein. The noncoding sequences have been considered junk, but perhaps it actually helps to explain organisms' complexity.

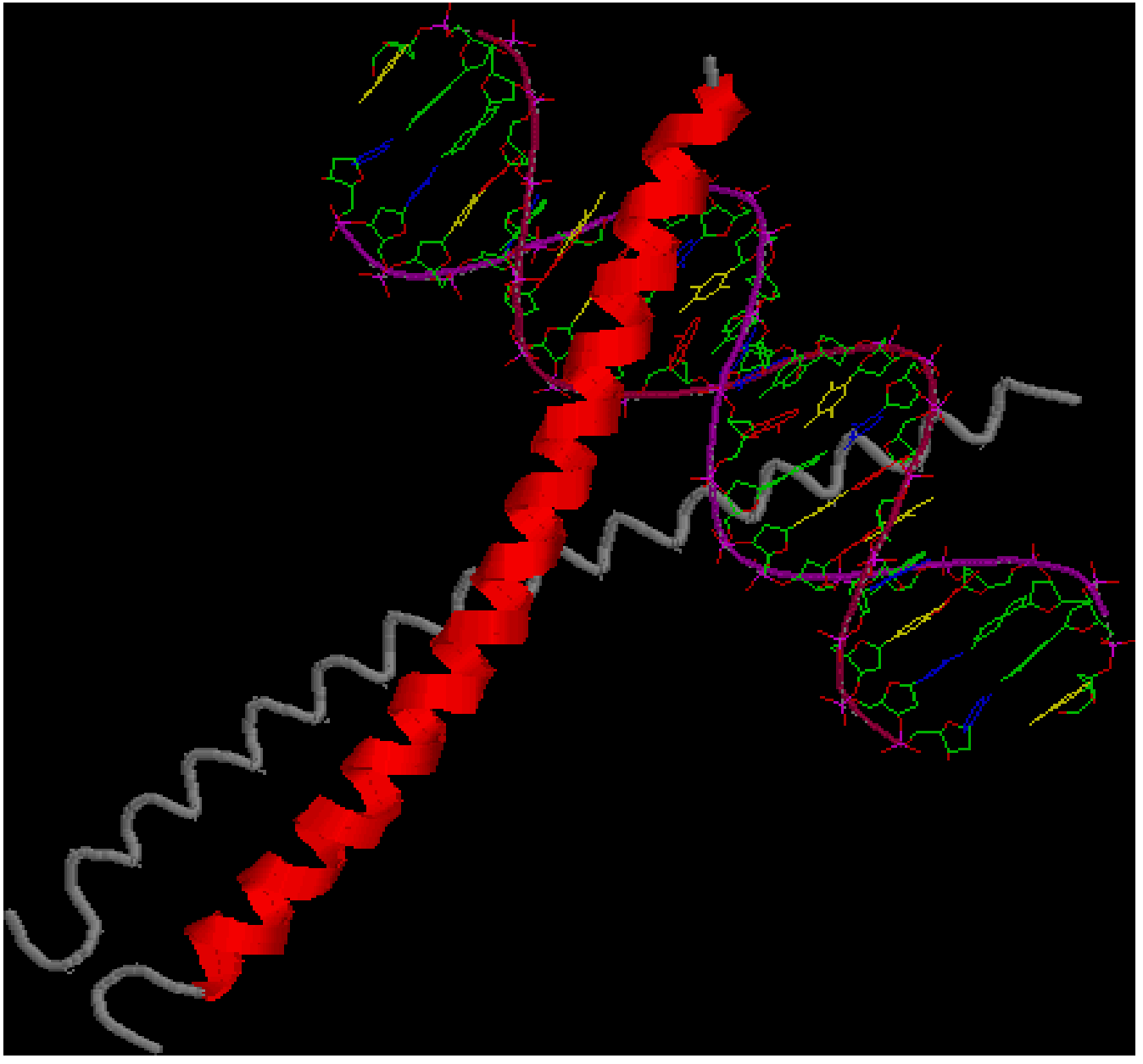
Spójrzmy na inicjację transkrypcji



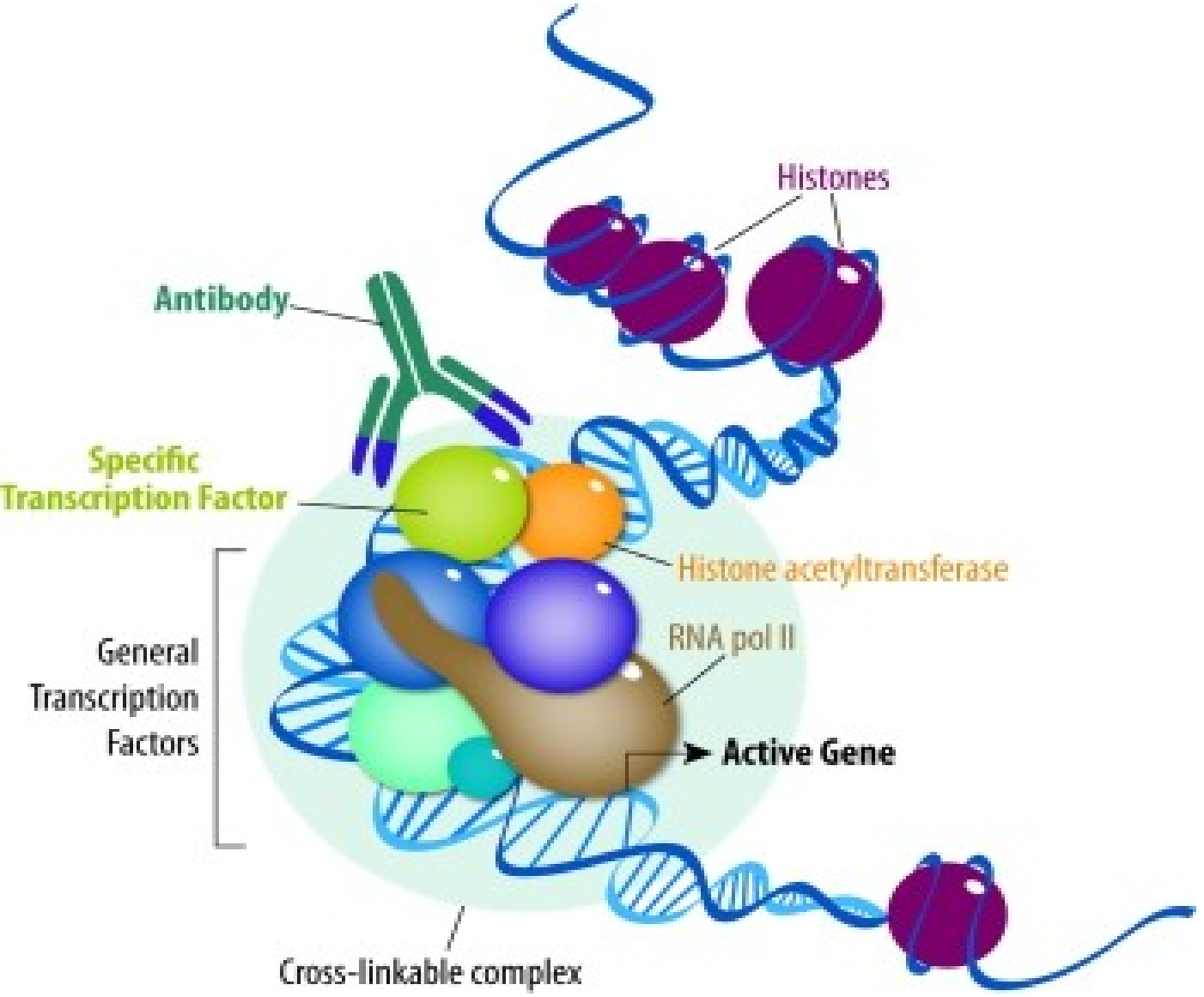
Rejon promotorowy



Wiązanie czynników transkrypcyjnych



Szukanie miejsc wiązania - ChIP



Reprezentacja miejsc wiązania

- Słowa konsensusowe
- kody zdegenerowane (np. IUPAC)
- Wyrażenia regularne
- Position specific frequency Matrices (macierze prawdopodobieństwa)
- Position specific Weight Matrices (log-odds)
- Modele Markowa wyższego rzędu
- Modele z zależnymi pozycjami (np. Sieci Bayesowskie)

Macierze ocen

- Zamiast uliniowienia miejsc wiązania możemy użyć macierzy zliczeń, zapominając o zależnościach pomiędzy pozycjami

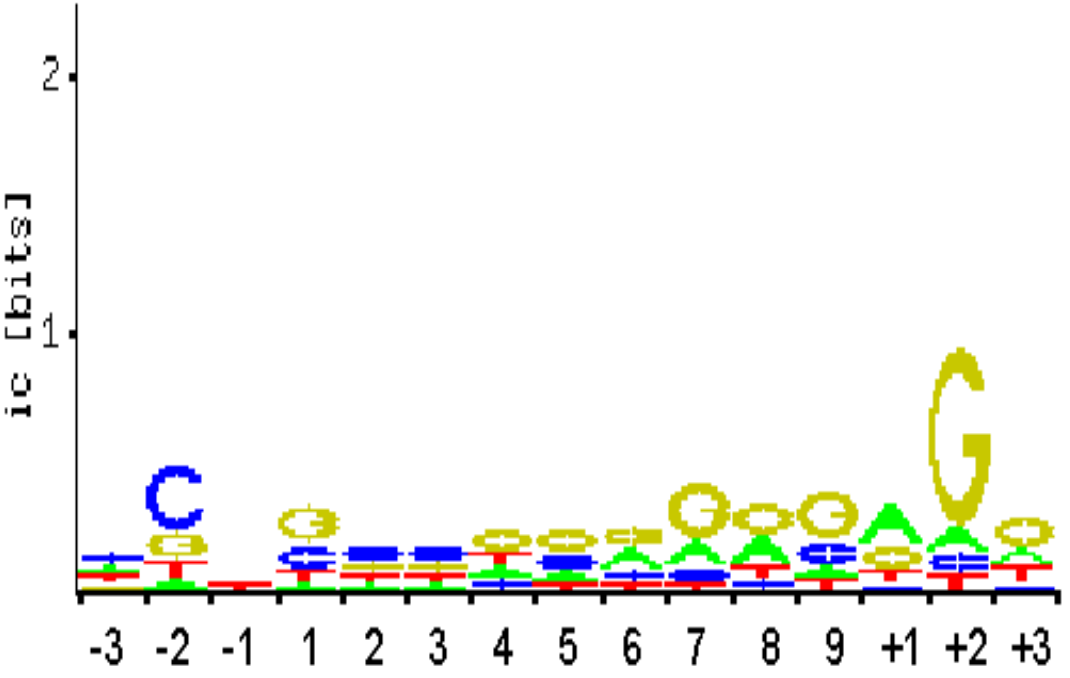
A	3	2	0	12	0	0	0	0	1	3
C	5	2	12	0	12	0	1	0	2	1
G	3	7	0	0	0	12	0	7	5	4
T	1	1	0	0	0	0	11	5	4	4

- Jeśli znormalizujemy kolumny do 1, mówimy o macierzy prawdopodobieństw

Model tła i logarytm szans(log-odds)

- Aby oszacować szanse związania słowa z motywem, musimy zdefiniować model “tła”, czyli losowej sekwencji DNA
- Kiedy mamy prawdopodobieństwo uzyskania nukleotydu i w losowej sekwencji (b_i) możemy zdefiniować logarytm szans $\log(\frac{p_i}{b_i})$
- Aby uniknąć logarytmu z 0, musimy stosować poprawkę Laplace'a (pseudozliczenia)
- Suma ocen z motywu to suma logarytmów

Możemy reprezentować macierze PWM jako logo sekwencji



A:	21	7	21	9	13	12	18	15	34	23	35	12	46	10	20
C:	39	58	21	32	40	37	14	27	18	14	10	23	7	7	10
G:	19	22	22	46	24	30	47	44	37	55	42	53	31	77	49
T:	20	12	33	12	21	19	19	12	9	6	10	9	14	5	19

- Log-odds na pozycji to wzajemna entropia modelu tła I modelu (rozkładu) motywu
- Jeśli model tła jest jednorodny (GC=50%) to jest to równoważne entropii rozkładu motywu

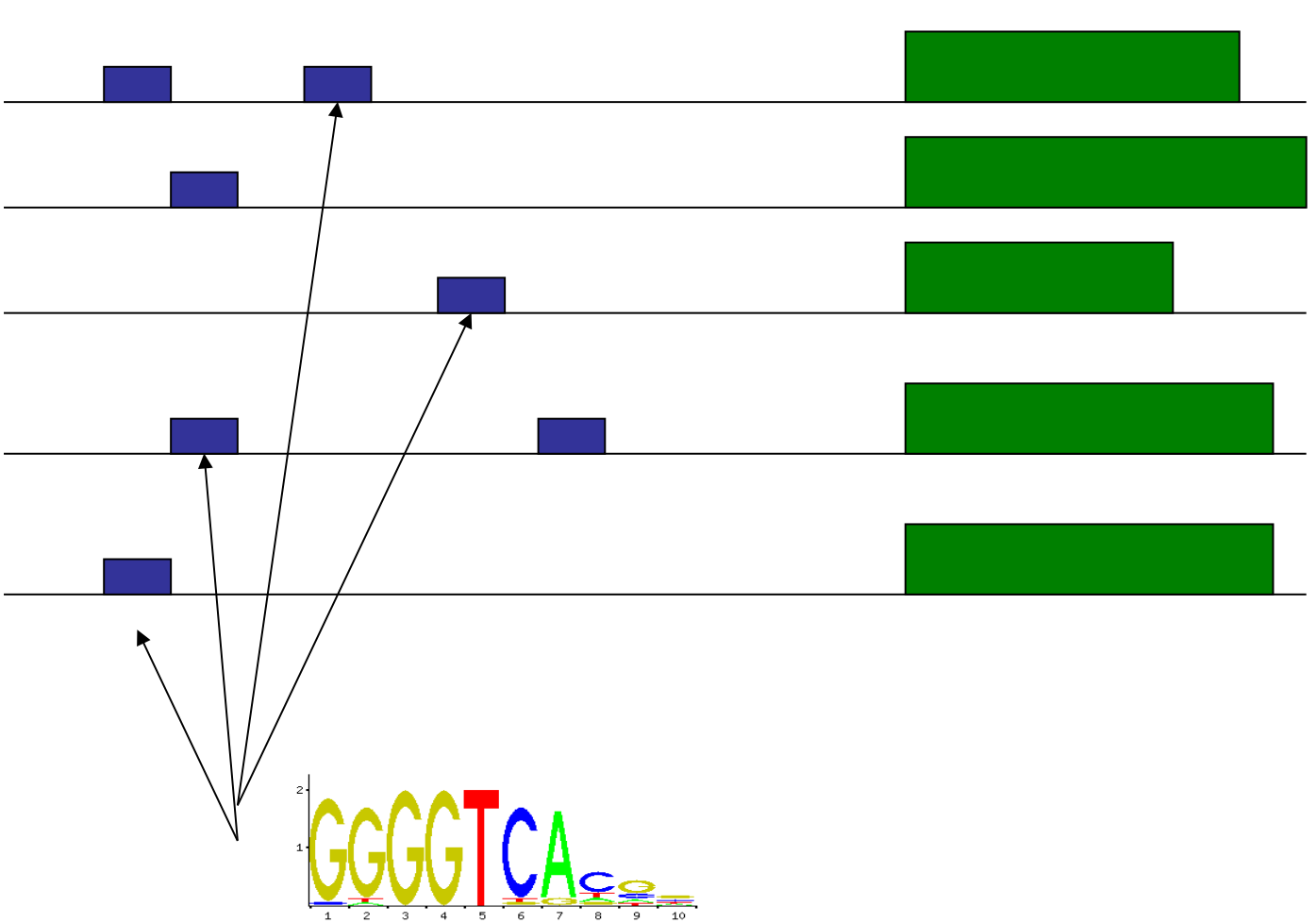
Wyszukiwanie miejsc wiązania w genomie

- Dla dowolnego słowa o długości równej długości motywu możemy obliczyć jego ocenę:

$$\sum_i \sum_j p_{ij} log_2 \frac{p_{ij}}{b_j}$$

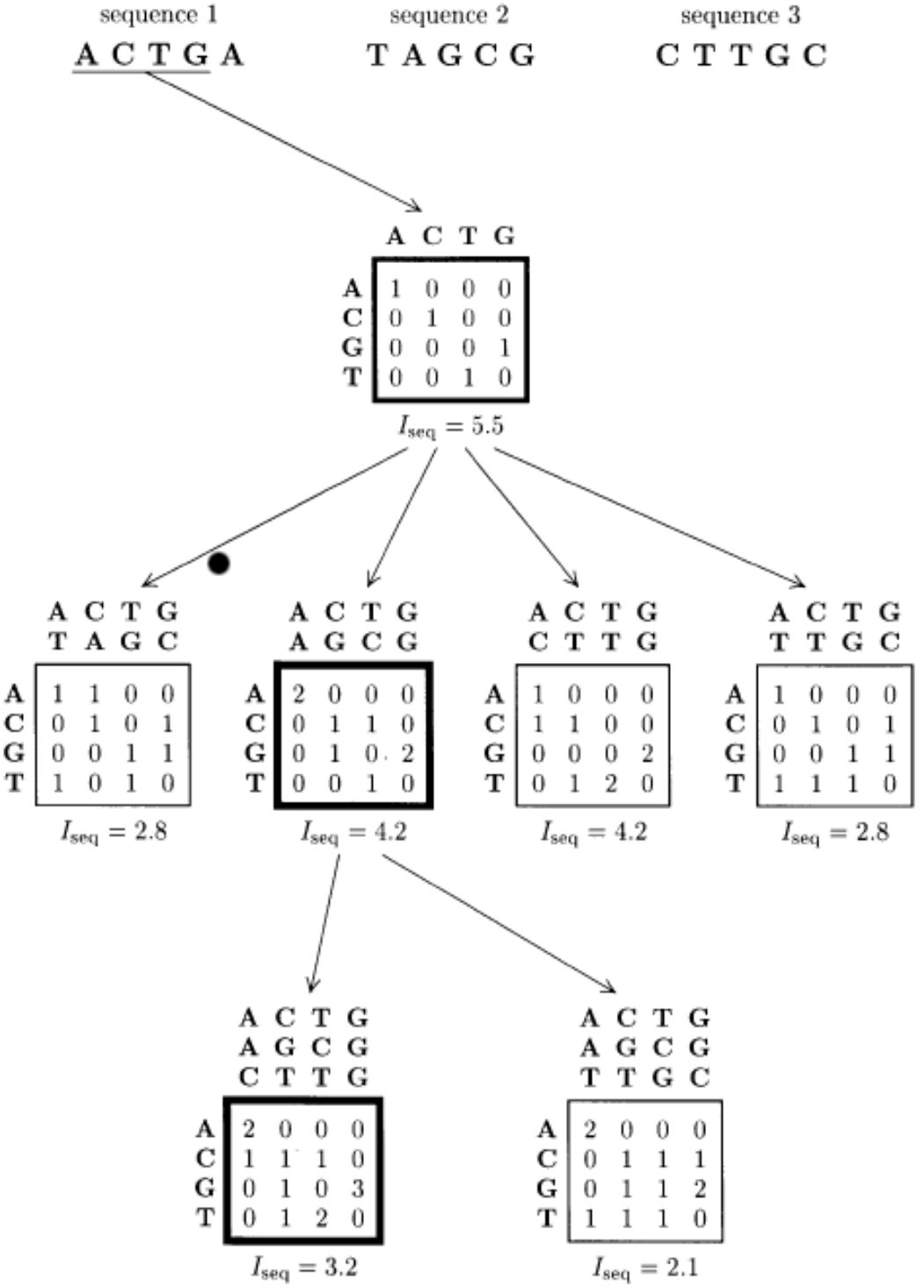
- Wysoka ocena daje większe szanse, że dane słowo odpowiada motywowi
- Możemy obliczań prawdopodobieństwo błędu na podstawie rozkładu ocen dla motywu
- Motywy możemy znaleźć w bazach danych TRANSFAC, JASPAR, HOCOMOCO I innnych

Znajdowanie nowych motywów

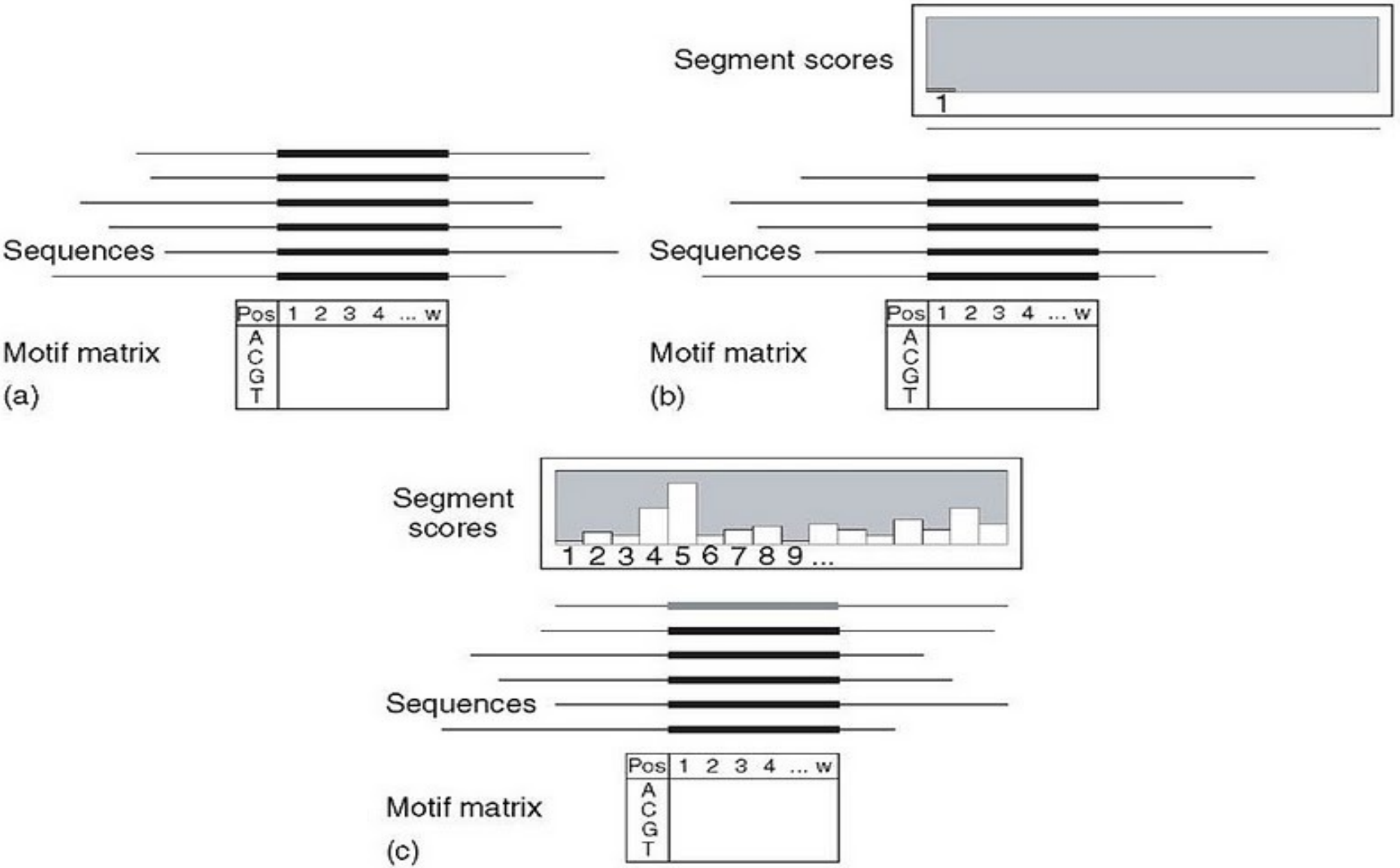


Metoda zachłanna

- Zależy od kolejności sekwencji
- Warto startwoać z różnych początkowych ustawień w uliniowieniu



Próbnik Gibbs'a



Motif Elucidation by Expectation Maximization (MEME)

S	unaligned set of sequences (training data) S_1, S_2, …, S_i, …, S_n each of length L
W	width of motif
Z	matrix of probabilities that the motif starts in position j in S_i
ρ	matrix representing the probability of character c in column k (the character c will be A, C, G, or T for DNA sequences or one of the 20 protein characters)
ε	epsilon value

```
1.  EM (S, W) {
2.      choose starting point and initial value for ρ
3.      do {
4.          re-estimate Z from ρ  //the estimation step
5.          re-estimate ρ from Z  //the maximization step
6.      } until (change in ρ < ε)
7.      return ρ , Z
8.  }
```