

# Przewidywanie miejsc wiązania nukleosomów w genomie drożdży

Aleksander Jankowski

Uniwersytet Warszawski  
Wydział Matematyki, Informatyki i Mechaniki

8 października 2009

Promotorzy:

Jerzy Tiuryn  
Uniwersytet Warszawski



Shyam Prabhakar  
Genome Institute of Singapore

# Outline

- 1 Wprowadzenie
- 2 Model wiązania nukleosomów
- 3 Model termodynamiczny
- 4 Wyniki
- 5 Podsumowanie

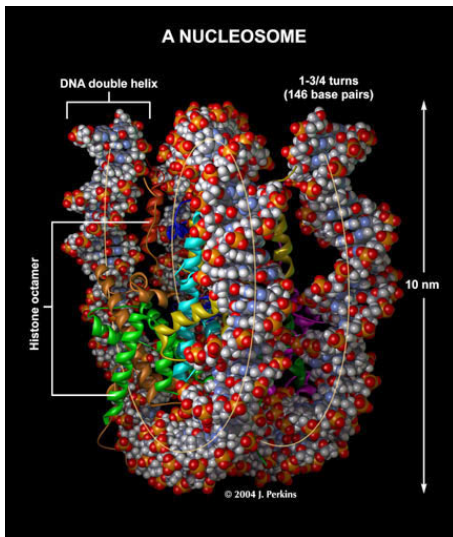
# Outline

- 1 Wprowadzenie
- 2 Model wiązania nukleosomów
- 3 Model termodynamiczny
- 4 Wyniki
- 5 Podsumowanie



# Organizacja nukleosomów

- Rdzeń każdego nukleosomu jest kompleksem ośmiu białek histonowych.
- Nawija się nań fragment podwójnej helisy DNA długości 147 par zasad (bp).
- Symetryczna struktura rdzenia wymusza bardzo regularny i zasadniczo powtarzający się układ nukleosomów.
- Nukleosomy występują na przemian z „łącznikowymi” fragmentami DNA długości 5-80 bp, zazwyczaj około 50 bp. Struktura zbudowana z nukleosomów i łączników bywa nazywana „koralami na nici”.
- Organizacja nukleosomów jest ściśle związana z regulacją transkrypcji genów, pozwalając na łatwiejszy dostęp do regionów pełniących funkcje regulatorowe.
- Struktura nukleosomów jest niemal identyczna u wszystkich eukariotów żyjących na Ziemi.



Źródło: <http://bio.research.ucsc.edu/people/boeger/ResearchInterests.htm>

# Podstawowa referencja i źródło danych

nature

Vol 458 | 19 March 2009 | doi:10.1038/nature07667

## LETTERS

### The DNA-encoded nucleosome organization of a eukaryotic genome

Noam Kaplan<sup>1\*</sup>, Irene K. Moore<sup>3\*</sup>, Yvonne Fondufe-Mittendorf<sup>3</sup>, Andrea J. Gossett<sup>4</sup>, Desiree Tillo<sup>5</sup>, Yair Field<sup>1</sup>, Emily M. LeProust<sup>6</sup>, Timothy R. Hughes<sup>5,7,8</sup>, Jason D. Lieb<sup>4</sup>, Jonathan Widom<sup>3</sup> & Eran Segal<sup>1,2</sup>

Nucleosome organization is critical for gene regulation<sup>1</sup>. In living cells this organization is determined by multiple factors, including the action of chromatin remodellers<sup>2</sup>, competition with site-specific DNA-binding proteins<sup>3</sup>, and the DNA sequence preferences of the nucleosomes themselves<sup>4,5</sup>. However, it has been difficult to estimate the relative importance of each of these mechanisms *in vivo*<sup>7,9–11</sup>, because *in vivo* nucleosome maps reflect the combined action of all influencing factors. Here we determine the importance of nucleosome DNA sequence preferences experimentally by measuring the genome-wide occupancy of nucleosomes assembled on purified yeast genomic DNA. The resulting map, in which nucleosome occupancy is governed only by the intrinsic sequence preferences of nucleosomes, is similar to *in vivo*

pair and the genome-wide average coverage per base pair (see Methods).

The nucleosome organizations of the *in vitro* and *in vivo* maps are notably similar, although not identical (Fig. 1), with a correlation of 0.74 between the nucleosome occupancy per base pair (Fig. 2a). On the scale of individual nucleosomes, the *in vitro* data separate regions that are enriched in nucleosomes *in vivo* from regions depleted of nucleosomes with high accuracy (Supplementary Fig. 1). Similarly, we found a significant correspondence between the positions of stable nucleosomes in the two maps (Supplementary Fig. 2). This high degree of similarity between the maps indicates that nucleosome sequence preferences have a dominant role in determining *in vivo* nucleosome organization.

# Dostępne dane

- W roku 2009, Kaplan i wsp. uzyskali pierwsze mapy pokrycia całego genomu drożdży nukleosomami *in vitro* i *in vivo*.
- Celem ich eksperymentu było zbadanie pokrycia nukleosomami *in vitro*, zależnego tylko od powinowactwa nukleosomów do sekwencji DNA.
- Wiązanie nukleosomów *in vitro* nie jest zaburzone przez konkurencję czynników transkrypcyjnych i innych czynników wiążących się z DNA.
- Do moich obliczeń używałem mapy pokrycia nukleosomami *in vitro*, uzyskanej z 9.3M odczytów sekwencji związanych z nukleosomami, dających się jednoznacznie zmapować na genom drożdży. Wśród odczytów było 5.3M różnych fragmentów sekwencji.



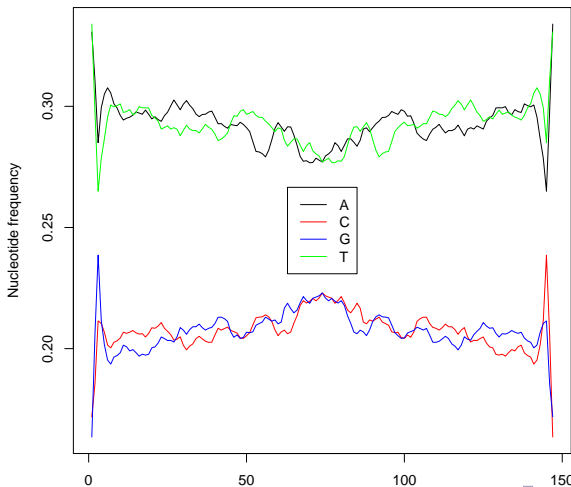
# Outline

- 1 Wprowadzenie
- 2 Model wiązania nukleosomów**
- 3 Model termodynamiczny
- 4 Wyniki
- 5 Podsumowanie

# Model wiązania nukleosomów

- Budujemy model probabilistyczny, który opíše powinowactwo nukleosomu do różnych fragmentów sekwencji DNA.
- Naszym celem jest przypisanie każdej sekwencji genomowej o długości 147 pewnej liczby rzeczywistej, opisującej dopasowanie tej sekwencji do nukleosomu.
- *Składowa pozycyjna* opisuje powinowactwo pary nukleotydów na ustalonej pozycji na nukleosomie do nukleosomu.
- *Składowa niezależna od pozycji* opisuje uśrednione powinowactwo sekwencji długości 5 do nukleosomu.
- Nazewnictwo:
  - sekwencja długości  $k$  =  $k$ -mer  
(np. sekwencja długości 5 = 5-mer)
  - para nukleotydów = 2-mer = dinukleotydy

# Powinowactwo poszczególnych nukleotydów do nukleosomu zależy od pozycji na nukleosomie



# Składowa pozycyjna

- Załóżmy, że mamy ustalone wagi  $N_i(pq)$  dla każdego  $1 \leq i \leq 146$ ,  $p, q \in \Sigma = \{A, C, G, T\}$ .
- Będziemy interpretować  $N_i(pq)$  jako prawdopodobieństwo napotkania dinukleotydu  $pq$  na  $i$ -tej pozycji nukleosomu. W szczególności  $N_i(\cdot)$  jest rozkładem prawdopodobieństwa.
- Możemy uogólnić definicję  $N_i$  aby uwzględnić pojedyncze nukleotydy.
- Dla  $1 \leq i \leq 146$  i  $p \in \Sigma$ , kładziemy  $N_i(p) = \sum_{q \in \Sigma} N_i(pq)$ .
- Załóżmy, że  $S$  jest sekwencją genomową długości 147. Wówczas definiujemy składową pozycyjną  $\mathcal{N}$  następująco:

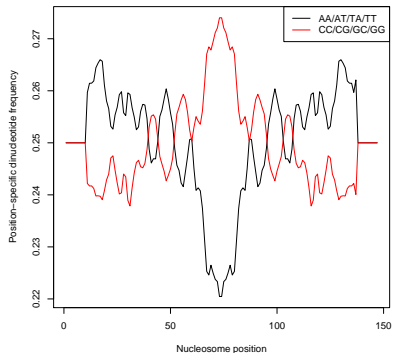
$$\mathcal{N}(S) = N_1(S_1) \cdot \prod_{i=1}^{146} \frac{N_i(S_i S_{i+1})}{N_i(S_i)} = N_1(S_1 S_2) \cdot \prod_{i=2}^{146} \frac{N_i(S_i S_{i+1})}{N_i(S_i)}.$$

# Składowa pozycyjna

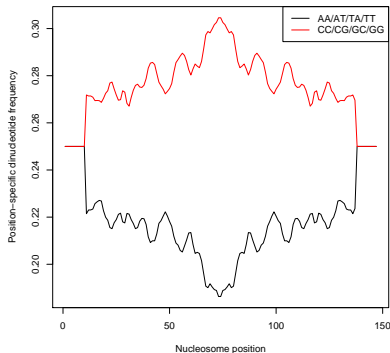
- Niech  $i$  będzie pozycją na nukleosomie, zaś  $p, q$  nukleotydami.
- Wagi  $N_i(pq)$  zostały wyliczone na podstawie odczytów sekwencji związanych z nukleosomami *in vitro*.
- Aby zapobiec efektom brzegowym w miejscach cięcia przez nukleazę mikrokokalną, używamy tylko środkowych 127 bp z sekwencji związanych z nukleosomami.
- Wbudowujemy symetrię w model:  $N_i(pq) = N_{146-i}(pq)$ .
- Rozważamy trzy warianty:
  - Double-normalised variant (**D**):  $n_i(pq) = \frac{\#_i(pq)}{\sum_{j=1}^{146} \#_j(pq)}$ .
  - Relative variant (**R**):  $n_i(pq) = \frac{\#_i(pq)}{\text{Overall}(pq)}$ , gdzie Overall( $\cdot$ ) jest rozkładem prawdopodobieństwa dinukleotydów w całym genomie.
  - Null variant (**Z**).

# Warianty składowej pozycyjnej

## Double-normalised variant (D)



## Relative variant (R)



# Składowa niezależna od pozycji

- Załóżmy, że mamy ustalone wagi  $L(p_1 \dots p_5)$  dla każdego 5-meru  $p_1, \dots, p_5 \in \Sigma$ .
- Będziemy interpretować  $L(p_1 \dots p_5)$  jako prawdopodobieństwo napotkania 5-meru  $p_1 \dots p_5$  w sekwencji związanej z nukleosomem. W szczególności  $L(\cdot)$  jest rozkładem prawdopodobieństwa.
- Podobnie jak dla składowej pozycyjnej, możemy uogólnić definicję  $L$  aby uwzględnić 4-mery.
- Dla  $p_1, \dots, p_4 \in \Sigma$ , kładziemy  $L(p_1 \dots p_4) = \sum_{q \in \Sigma} L(p_1 \dots p_4 q)$ .
- Załóżmy, że  $S$  jest sekwencją genomową długości 147. Wówczas definiujemy składową niezależną od pozycji  $\mathcal{L}$  następująco:

$$\mathcal{L}(S) = L(S_1 \dots S_5) \cdot \prod_{i=2}^{143} \frac{L(S_i \dots S_{i+4})}{L(S_i \dots S_{i+3})}$$

# Składowa niezależna od pozycji

- Niech  $p_1 \dots p_5$  będzie 5-merem (sekwencją nukleotydów długości 5).
- Wagi  $L(p_1 \dots p_5)$  zostały wyliczone na podstawie odczytów sekwencji związanych z nukleosomami *in vitro*.
- Aby zapobiec efektom brzegowym w miejscach cięcia przez nukleazę mikrokokalną, używamy tylko środkowych 127 bp z sekwencji związanych z nukleosomami.
- Rozważamy trzy warianty:

- Natural (**N**):  $L(p_1 \dots p_5) = \frac{\#(p_1 \dots p_5)}{\sum_{q_1, \dots, q_5 \in \Sigma} \#(q_1 \dots q_5)}$ .
- Relative (**R**):

$$L(p_1 \dots p_5) = \frac{\frac{\text{Overall}(p_1 \dots p_5)}{\#(p_1 \dots p_5)}}{\sum_{q_1, \dots, q_5 \in \Sigma} \frac{\text{Overall}(q_1 \dots q_5)}{\#(q_1 \dots q_5)}}$$

- Null (**Z**):  $L(p_1 \dots p_5) = \left(\frac{1}{4}\right)^5$ .



# Łączna miara dopasowania

- Załóżmy, że  $S$  jest sekwencją genomową długości 147.
- Łączna miara dopasowania jest wyliczana jako

$$\text{Score}(S) = \ln \frac{\mathcal{L}(S)}{\mathcal{N}(S)}.$$

# Outline

- 1 Wprowadzenie
- 2 Model wiązania nukleosomów
- 3 Model termodynamiczny**
- 4 Wyniki
- 5 Podsumowanie

# Termodynamiczny model wiązania nukleosomów

- Załóżmy teraz, że  $S$  jest sekwencją genomową całego chromosomu.
- Niech  $\mathcal{C}$  będzie przestrzenią wszystkich dopuszczalnych konfiguracji nukleosomów na sekwencji  $S$ . Dopuszczalną konfiguracją jest taki zbiór nukleosomów długości 147 bp (reprezentowanych przez ich pozycje początkowe na  $S$ ), że żadne dwa nukleosomy na siebie nie zachodzą.

# Termodynamiczny model wiązania nukleosomów

- Dla każdej konfiguracji  $c \in \mathcal{C}$ , składającej się z  $k$  nukleosomów o pozycjach początkowych  $c[1], \dots, c[k]$ , przypisujemy statystyczną wagę

$$W_c[S] = \prod_{i=1}^k \tau \cdot \exp(\beta \cdot \text{Score}(S[c[i] \dots c[i] + 146])),$$

gdzie  $\tau$  i  $\beta$  są ustalonymi parametrami, zaś  $S[k \dots l]$  oznacza podsekwencję  $S$  od pozycji  $k$  do pozycji  $l$  włącznie.

- Parametr  $\tau$  można interpretować jako stężenie nukleosomów, a  $\beta$  jako odwrotność temperatury. Przyjmując rozkład Boltzmann na  $\mathcal{C}$ , możemy określić prawdopodobieństwo każdej konfiguracji  $c \in \mathcal{C}$ :

$$P(c|S) = \frac{W_c[S]}{\sum_{c' \in \mathcal{C}} W_{c'}[S]}.$$

# Termodynamiczny model wiązania nukleosomów

- Znalezienie konfiguracji  $c \in \mathcal{C}$  maksymalizującej  $P(c|S)$  jest bardzo kłopotliwe, ze względu na wykładniczy koszt przeszukiwania całej przestrzeni kombinacji  $\mathcal{C}$ .
- Ponadto, przywiązywanie się do najbardziej prawdopodobnej konfiguracji może być zwodnicze, gdyż można zaniedbać podzbiór podobnych konfiguracji o większym łącznym prawdopodobieństwie.
- Zastosujemy inne podejście. naszym celem będzie policzenie, dla każdej pozycji na  $S$ , średniego pokrycia tej pozycji nukleosomami, rozumianego jako prawdopodobieństwo, że dana pozycja jest pokryta przez jakiś nukleosom.
- Aby to osiągnąć, użyjemy programowania dynamicznego, aby dla każdej pozycji na  $S$  policzyć prawdopodobieństwo, że jakiś nukleosom zaczyna się właśnie na tej pozycji.

# Kroczymy naprzód...

Wyliczamy ciąg  $F_1, \dots, F_M$ , gdzie  $M = |S|$ . Liczba  $F_i$  będzie sumą statystycznych wag wszystkich dopuszczalnych konfiguracji na podsekwencji  $S_1, \dots, S_i$ :

$$F_0 := 1 \quad (\text{dla pełności}),$$

$$F_i := F_{i-1} \quad \text{dla } 1 \leq i \leq 146,$$

$$F_i := F_{i-1} + F_{i-147} \cdot \tau \cdot \exp(\beta \cdot \text{Score}(S[i-146 \dots i])) \\ \text{dla } 147 \leq i \leq M.$$

Ten wzór wyraża fakt, że każda konfiguracja  $c$  na podsekwencji  $S_1, \dots, S_i$  spełnia dokładnie jeden z poniższych warunków:

- 1  $c$  nie pokrywa pozycji  $S[i]$ , wobec czego można o niej myśleć jak o konfiguracji nukleosomów na  $S_1, \dots, S_{i-1}$
- 2  $c$  zawiera nukleosom pokrywający pozycje  $S[i-146 \dots i]$  (a na podsekwencji  $S_1, \dots, S_{i-147}$  może być cokolwiek).

## ...i wstecz

Wyliczamy ciąg  $R_1, \dots, R_M$ , gdzie  $M = |S|$ . Liczba  $R_i$  będzie sumą statystycznych wag wszystkich dopuszczalnych konfiguracji na podsekwencji  $S_i, \dots, S_M$ :

$$\begin{aligned} R_{M+1} &:= 1 \quad (\text{dla pełności}), \\ R_i &:= R_{i+1} \quad \text{dla } M - 145 \leq i \leq M, \\ R_i &:= R_{i+1} + R_{i+147} \cdot \tau \cdot \exp(\beta \cdot \text{Score}(S[i \dots i + 146])) \\ &\quad \text{dla } 1 \leq i \leq M - 146. \end{aligned}$$

# Łączymy kroki naprzód z krokami wstecz

Wpierw zauważmy, że z definicji  $F_i$  i  $R_i$ ,

$$F_M = R_1 = \sum_{c \in \mathcal{C}} W_c[S].$$

Możemy teraz łatwo policzyć prawdopodobieństwo  $P_i[S]$  umieszczenia nukleosomu zaczynającego się na pozycji  $1 \leq i \leq M - 146$  sekwencji  $S$ :

$$P_i[S] = \frac{F_{i-1} \cdot \tau \cdot \exp(\beta \cdot \text{Score}(S[i \dots i + 146])) \cdot R_{i+147}}{R_1}.$$

Średnie pokrycie nukleosomami, rozumiane jako prawdopodobieństwo, że dana pozycja  $i$  jest pokryta przez jakiś nukleosom, może być teraz wyliczone następująco:

$$\sum_{k=0}^{146} P_{i-k}[S].$$



# Outline

- 1 Wprowadzenie
- 2 Model wiązania nukleosomów
- 3 Model termodynamiczny
- 4 Wyniki**
- 5 Podsumowanie

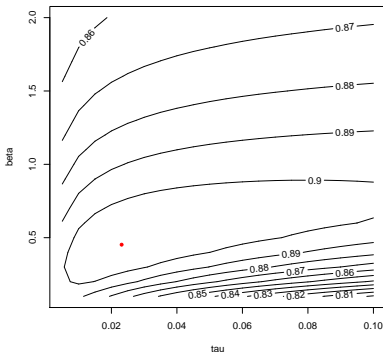
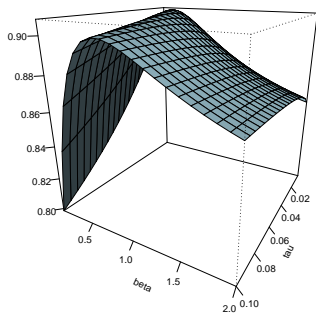
# Optymalne wartości $\tau$ i $\beta$ dla genomu drożdży

		Składowa pozycyjna ( $\mathcal{N}$ )		
		D	R	Z
Skt. nzl. od pozycji ( $\mathcal{L}$ )	N	$\tau = 0.0160$ $\beta = 0.3415$ korelacja <b>59.9%</b>	$\tau = 0.0354$ $\beta = 0.2956$ korelacja <b>71.7%</b>	$\tau = 0.0148$ $\beta = 0.3353$ korelacja <b>59.2%</b>
	R	$\tau = 0.0105$ $\beta = 0.7918$ korelacja <b>90.3%</b>	$\tau = 0.0231$ $\beta = 0.4524$ korelacja <b>90.8%</b>	$\tau = 0.0100$ $\beta = 0.7593$ korelacja <b>89.4%</b>
Z	$\tau = 8.16 \cdot 10^{-13}$ $\beta = 3.50 \cdot 10^{-6}$ korelacja <b>37.4%</b>	$\tau = 0.0306$ $\beta = 0.9160$ korelacja <b>87.5%</b>	—	

Optymalne wartości  $\tau$  i  $\beta$  dla całej nierepetytywnej części genomu drożdży (10.8M bp = 89.3% genomu).

Dla tych wartości  $\tau$  i  $\beta$  podana jest też korelacja Pearsona między przewidywanym i obserwowanym *in vitro* pokryciem nukleosomami.

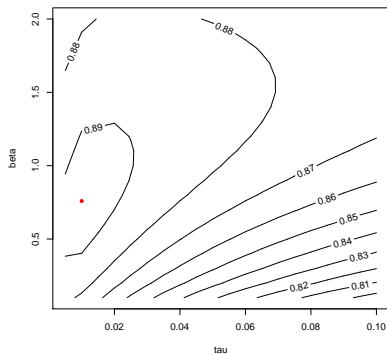
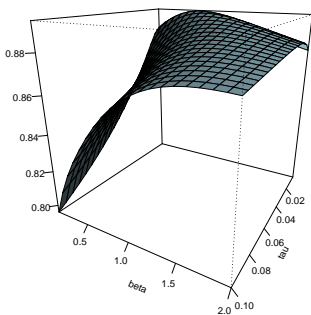
# Przewidywania pełnego modelu



Optymalne wartości:  $\tau = 0.0231$ ,  $\beta = 0.4524$ , korelacja 90.8%.

Dla porównania, najlepszym wynikiem uzyskanym przez Kaplana i wsp. jest 88.0%.

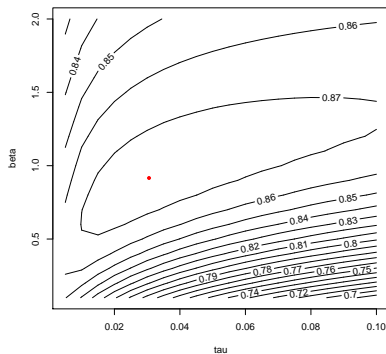
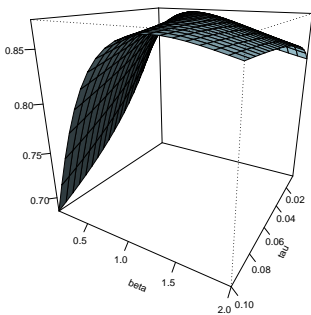
# Przewidywania samej składowej niezależnej od pozycji



Optymalne wartości:  $\tau = 0.0100$ ,  $\beta = 0.7593$ , korelacja 89.4%.

Dla porównania, wynikiem uzyskanym przez Kaplana i wsp. w analogicznym przypadku ( $P_L$ -only model) jest 87.6%.

# Przewidywania samej składowej pozycyjnej



Optymalne wartości:  $\tau = 0.0306$ ,  $\beta = 0.9160$ , korelacja 87.5%.

Dla porównania, wynikiem uzyskanym przez Kaplana i wsp. w analogicznym przypadku ( $P_N$ -only model) jest 82.0%.

# Outline

- 1 Wprowadzenie
- 2 Model wiązania nukleosomów
- 3 Model termodynamiczny
- 4 Wyniki
- 5 Podsumowanie**

# Podsumowanie

- Naszym celem jest stworzenie modelu dobrze opisującego wiązanie nukleosomów i czynników transkrypcyjnych.
- Celem pracy magisterskiej było powtórzenie wyników uzyskanych przez Kaplana i wsp. oraz zbadanie wpływu parametrów modelu na jego wierność.
- Opis użytych metod w oryginalnym artykule był bardzo niejasny, wobec czego testowałem różne warianty modelu wiązania nukleosomów.
- W najlepszych wariantach, uzyskaliśmy lepsze wyniki niż podane w oryginalnym artykule.

# Dalszy rozwój

- Całościowy model wiązania nukleosomów i czynników transkrypcyjnych.
- Więzy na długość łączników między kolejnymi nukleosomami (wymuszone przez interakcje między nimi w skali molekularnej).
- Uwzględnienie potwierdzonego doświadczalnie współdziałania między czynnikami transkrypcyjnymi.