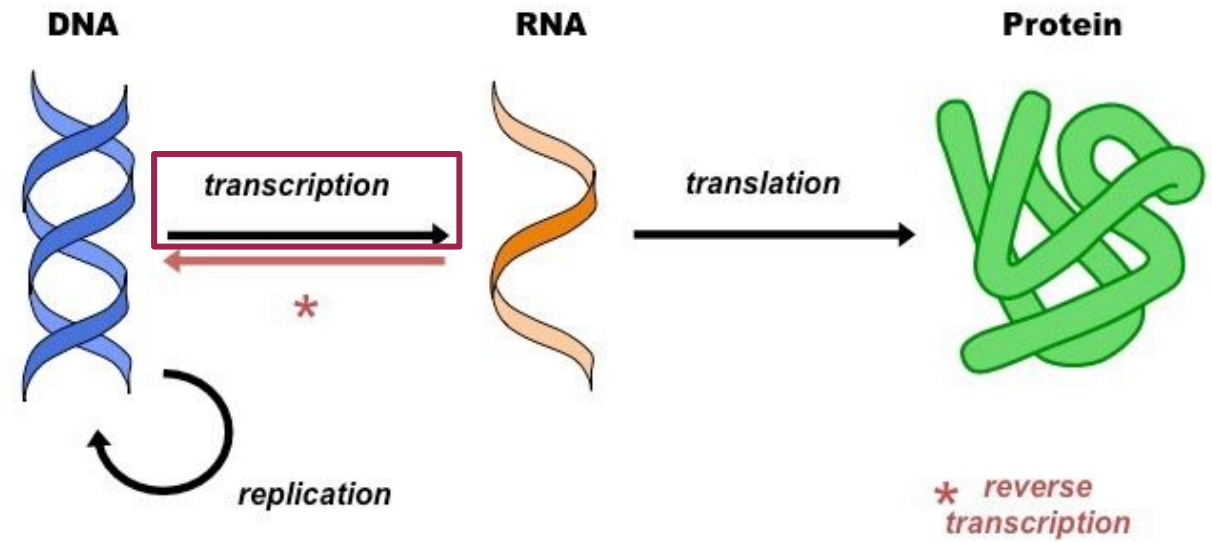
Base-resolution models of transcription factor binding reveal soft motif syntax

Žiga Avsec[1], Melanie Weilert[2], Avanti Shrikumar[3], Sabrina Krueger[2], Amr Alexandari[3], Khyati Dalal[2,5], Robin Fropf[2], Charles McAnany[2], Julien Gagneur[1], Anshul Kundaje[3,4]* and Julia Zeitlinger[2,5]*

[1] Department of Informatics, Technical University of Munich, Garching, Germany
[2] Stowers Institute for Medical Research, Kansas City, MO, USA
[3] Department of Computer Science, Stanford University, Stanford, CA, USA
[4] Department of Genetics, Stanford University, Stanford, CA, USA
[5] The University of Kansas Medical Center, Kansas City, KS, USA
* correspondence: akundaje@stanford.edu, jbz@stowers.org

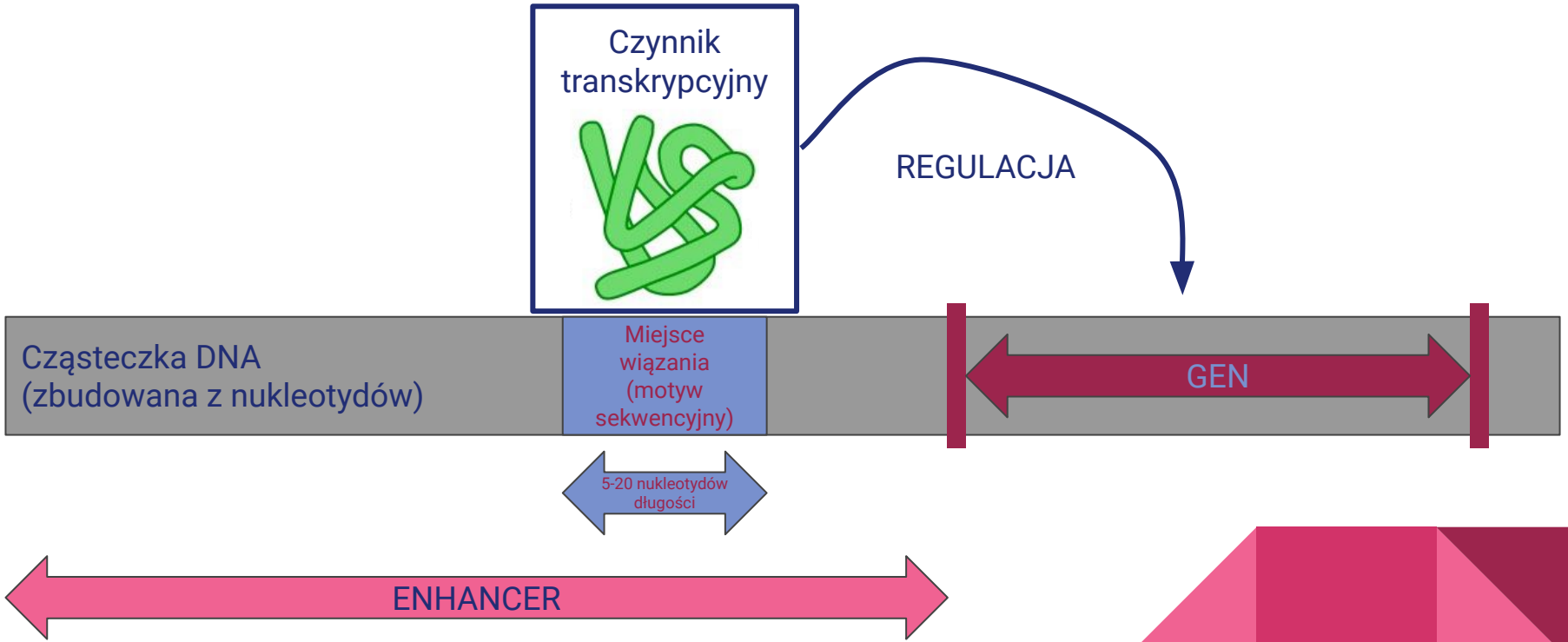
Autor prezentacji: Stanisław Antonowicz

Wstęp teoretyczny: centralny dogmat biologii molekularnej

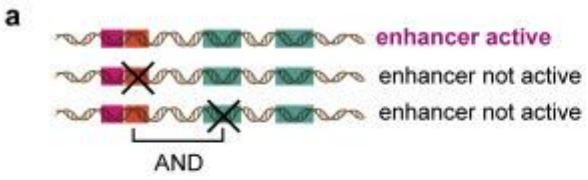


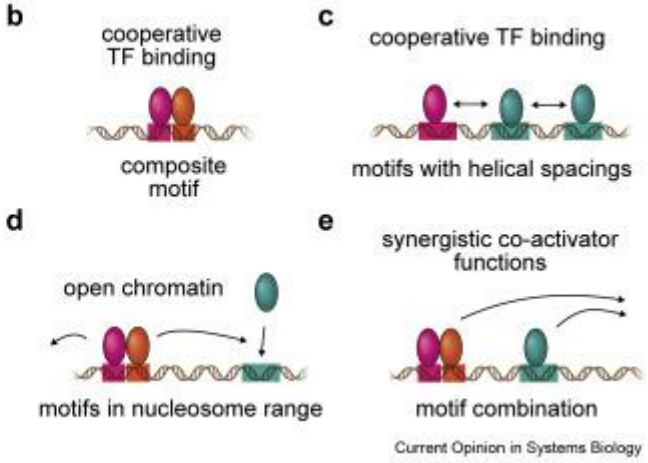
Źródło: https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/27-dna-replication-transcri/central-dogma.html

Wstęp teoretyczny: czynniki transkrypcyjne



Wstęp teoretyczny: składnia motywów





- Czynniki transkrypcyjne wiążąc się do genomu blisko siebie wchodzą ze sobą w interakcje.
- Rozłożenie motywów sekwencyjnych nazywamy składnią.
  - “Silna” składania to taka, gdzie motywy sekwencyjne oddziałują ze sobą w prosty sposób (np dwa motywy występują obok siebie). (b)
  - “Słaba” składnia to taka, gdzie ułożenie motywów na genomie jest bardziej skomplikowane bądź nieregularne. (c), (d), (e)

https://doi.org/10.1016/j.coisb.2020.08.002

Wstęp teoretyczny: znaczenie czynników transkrypcyjnych

- Czynniki transkrypcyjne regulują ekspresję genów.
- Dzięki temu komórki mogą:
  - Reagować na zmiany w ich otoczeniu (organizmu bądź komórki).
  - Różnicować się w różne typy komórek.
  - Kontrolować cykl komórkowy.
- Niektóre czynniki transkrypcyjne są (częściowo) odpowiedzialne za niekontrolowane różnicowanie się i wzrost komórek rakowych.
  - Na przykład czynnik transkrypcyjny Nanog występuje w komórkach raka piersi.
    - Lu, X., Mazur, S., Lin, T. *et al.* The pluripotency factor nanog promotes breast cancer tumorigenesis and metastasis. *Oncogene* 33, 2655–2664 (2014). https://doi.org/10.1038/onc.2013.209

Rozszyfrowanie znaczenia tytułu

Base-resolution models	Modele z dokładnością co do nukleotydu
of **transcription factor binding**	wiązanie czynników transkrypcyjnych
reveal **soft motif syntax**	“słaba” składnia motywów

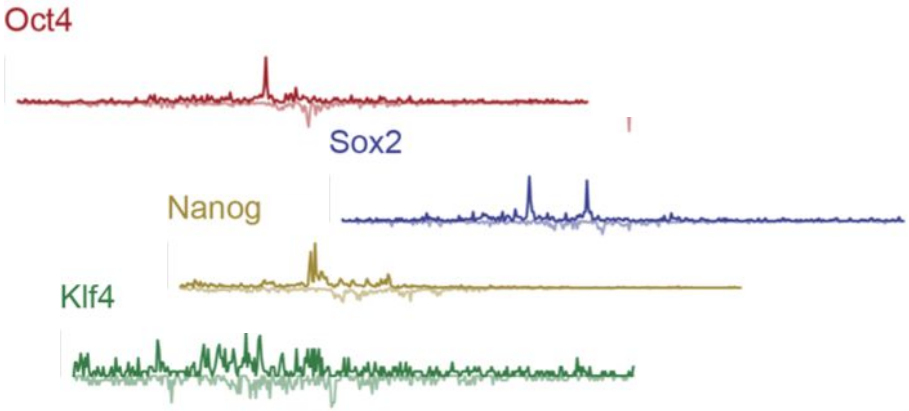
Architektura modelu – wejście i wyjście

Wejście:

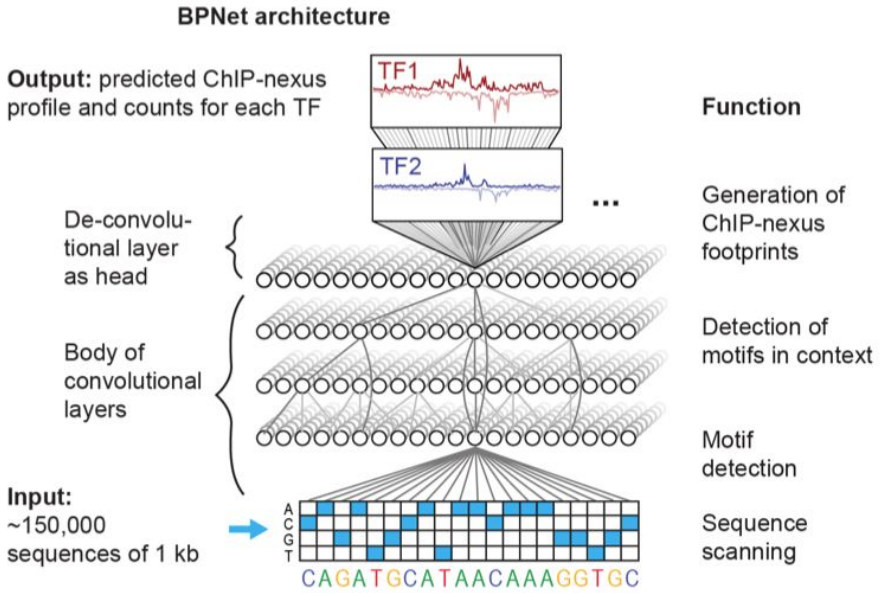
Sekwencje DNA w postaci macierzy wektorów “one-hot”.

Wyjście:

1. Profile częstotliwości wiązania czterech różnych czynników transkrypcyjnych w odcinku przed genem.
2. Ogólna liczba przyłączonych się cząsteczek danego czynnika transkrypcyjnego na danym odcinku.



Architektura modelu



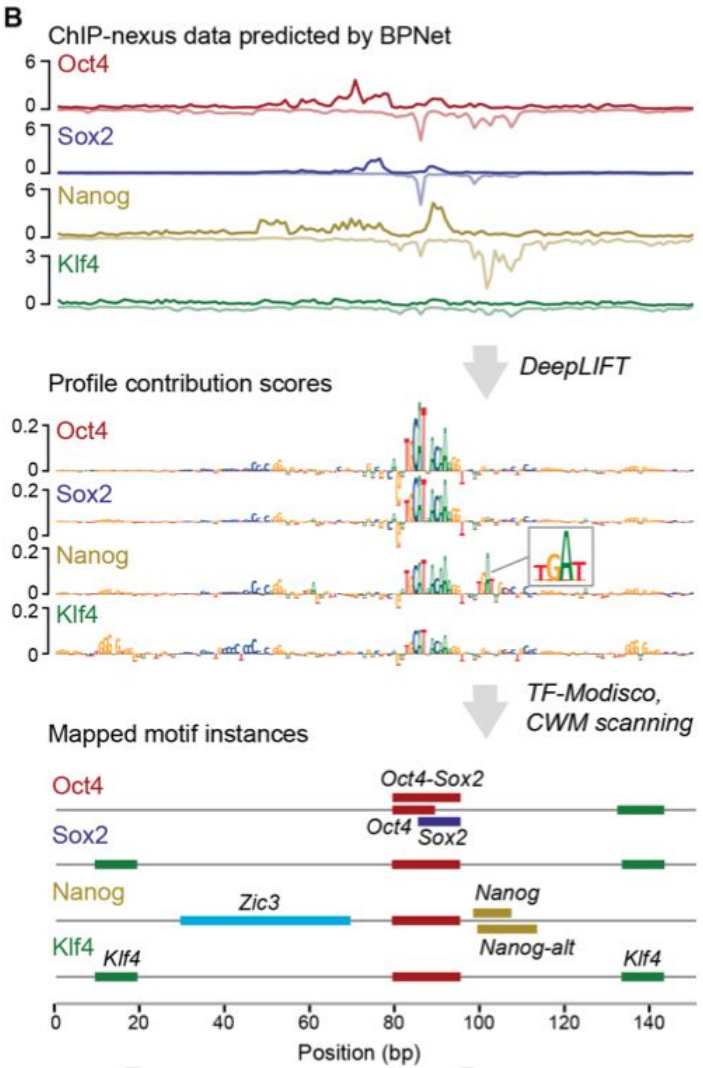
Kluczowe cechy modelu:

- Sieć konwolucyjna.
- Zastosowanie uczenia wielozadaniowego *(multitask learning)*.
- Rezygnacja z poolingu.
- Kształt oraz łączna “masa” profilu częstotliwości są przewidywane oddzielnie. Pozwala na to niestandardowa funkcja straty.
- Używa pomocniczych danych o warunkach przeprowadzenia eksperymentu, żeby uniknąć biasu wynikającego z czynników losowych występujących w trakcie tworzenia zbioru danych.

$$Loss = -\log p_{mult.}(\mathbf{k}^{obs}|\mathbf{p}^{pred}, n^{obs}) + \lambda(\log(1+n^{obs}) - \log(1+n^{pred}))^2$$

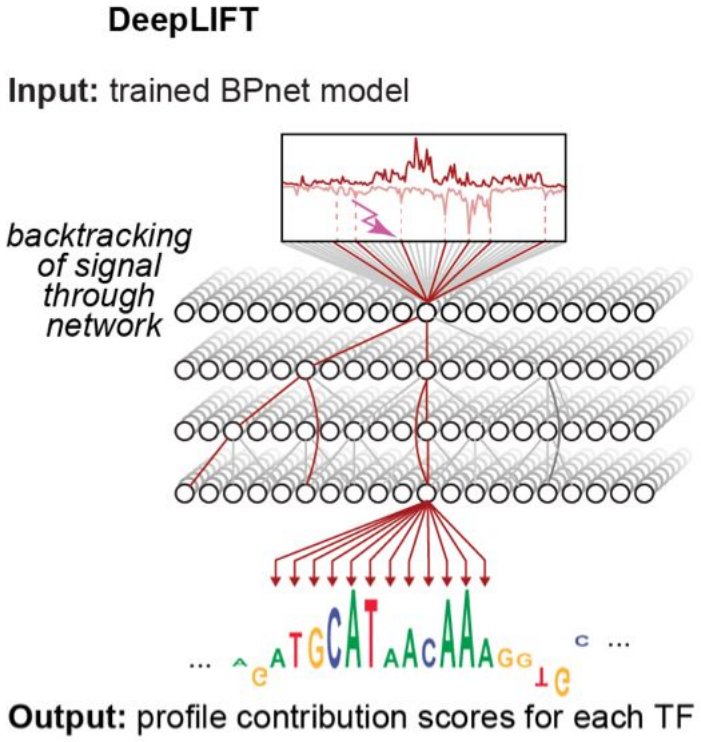
Pipeline interpretacyjny

1. Predykcja profilu wiązania czynników transkrypcyjnych dla danego fragmentu DNA.
2. Profile kontrybucji dla każdego nukleotydu z sekwencji wejściowej.
3. Identyfikacja występowania konkretnych motywów w sekwencji wejściowej.



Interpretacja modelu – DeepLIFT

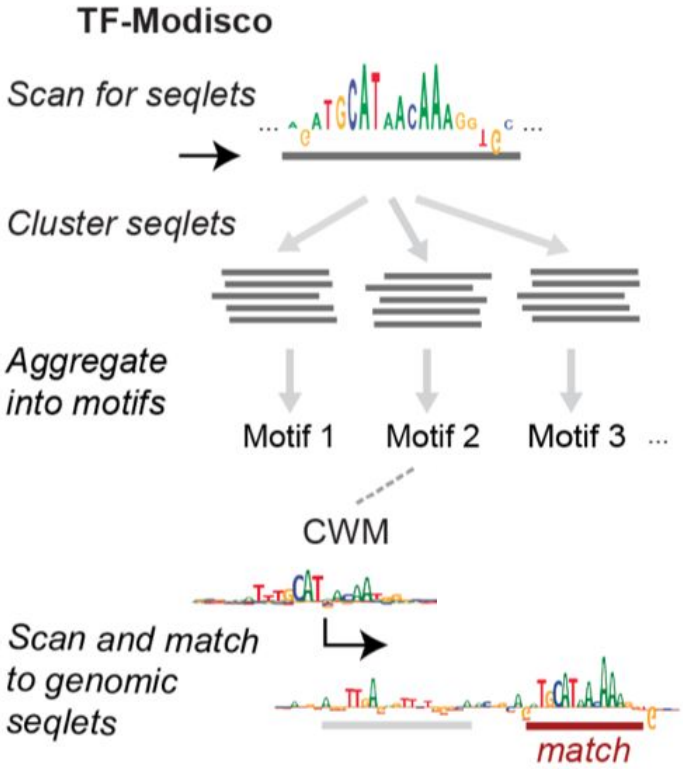
- Metoda gradientowa.
- Wyjaśniana jest różnica pomiędzy wyjściem sieci a wartością referencyjną (“neutralną”).
  - Dla MNIST może to być biały obraz, dla sekwencji DNA można użyć wielu permutacji sekwencji wejściowej i uśrednić wyniki.
  - Potrzebna jest wiedza domenowa, żeby zdefiniować dobrą referencję.
- Na drugim wykładzie ta metoda była wspomniana jako analogia do SHAPa.
- https://github.com/kundajelab/deeplift



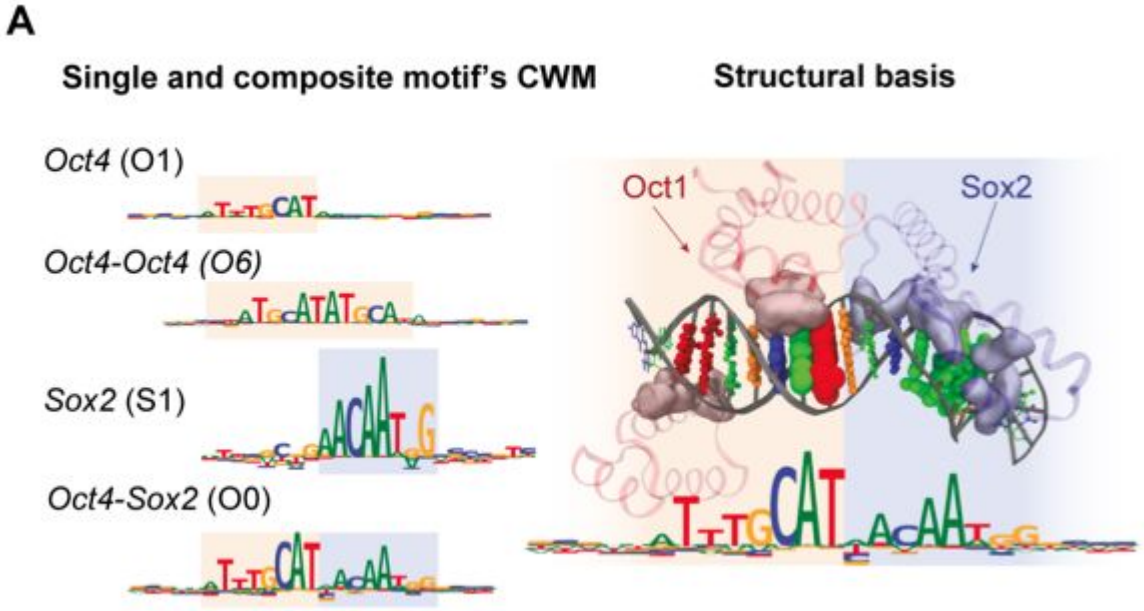
Interpretacja modelu – TF-MoDISco

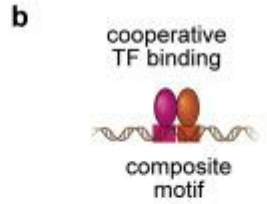
(TF MOtif Discovery from Importance SCOres)

- Motywy sekwencyjne mogą się nieznacznie różnić od siebie, a mimo to być miejscami wiązań czynników transkrypcyjnych.
- TF-MoDISco pozwala na znalezienie konkretnych, znanych motywów sekwencyjnych w analizowanych sekwencjach.
  - Krótkie sekwencje z dużymi kontrybucjami są klastrowane.
  - Następnie dla każdego klastra jest obliczana reprezentacja CWM (contribution weight matrix), przedstawiająca kontrybucję każdego nukleotydu na każdej pozycji w klastrze.
  - Wejściowe sekwencje są skanowane z użyciem wyżej wspomnianej reprezentacji, żeby zidentyfikować wystąpienia konkretnych motywów w sekwencji.
- https://github.com/kundajelab/tfmodisco



Wyniki – “silna” składania motywów

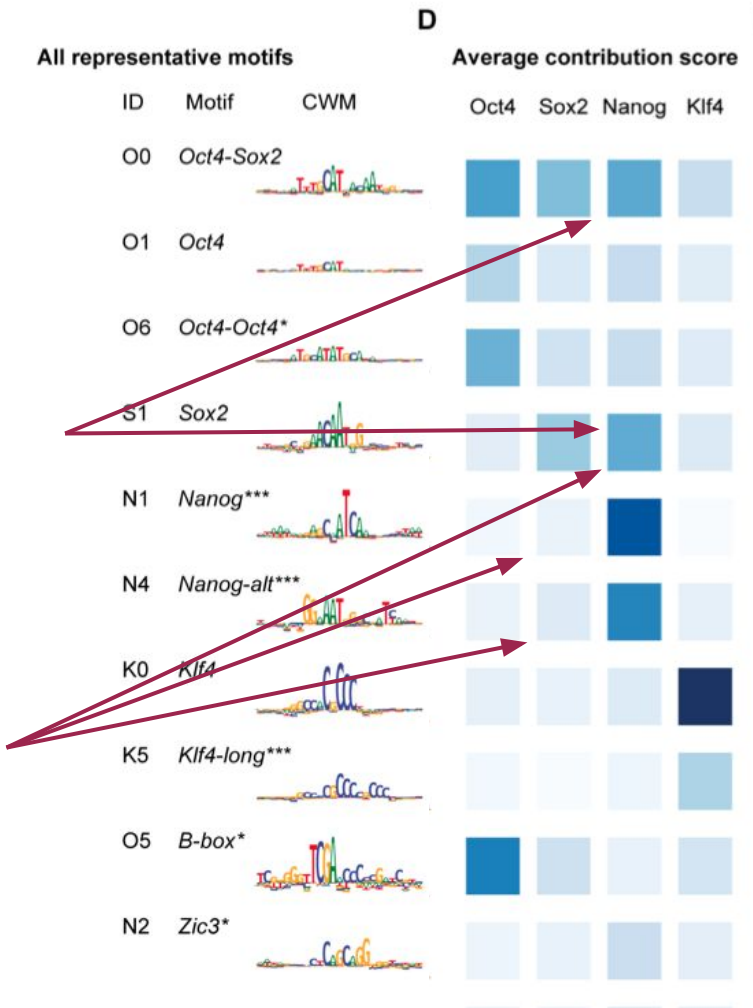




Wyniki – “słaba” składania motywów

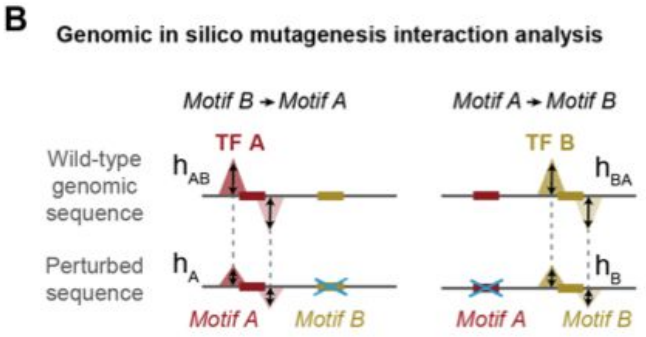
- Motywy *Oct4-Sox2* i *Sox2* wpływają na wiązanie czynnika transkrypcyjnego Nanog.

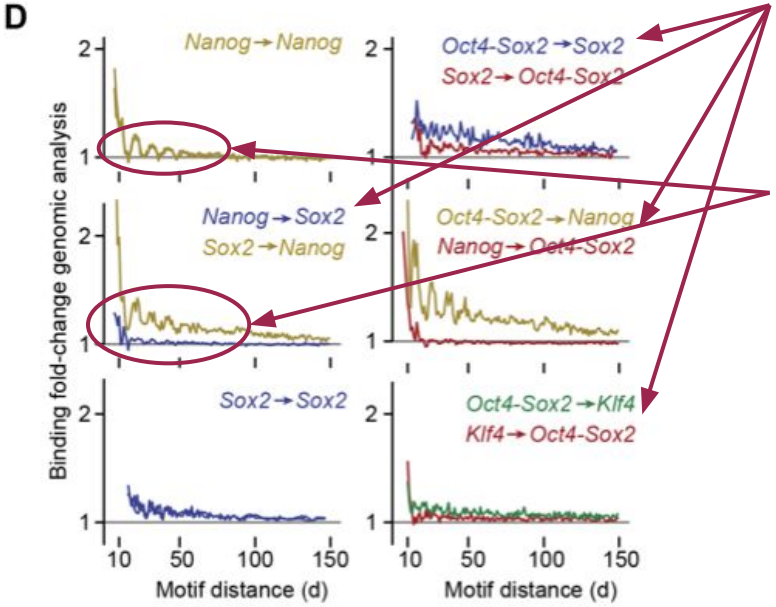
- Relacja jest jednokierunkowa – motyw *Sox2* wpływa na wiązanie Nanog, ale motywy *Nanog* nie wpływają na wiązanie Sox2.

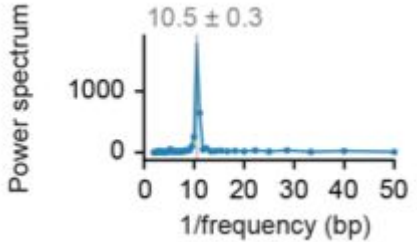


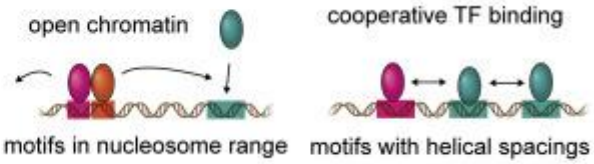
Wyniki – “słaba” składania motywów.

- Niesymetryczność wpływu motywów sekwencyjnych na wiązanie czynników transkrypcyjnych.
- Okresowy charakter wiązania Nanog.









Podsumowanie

- Konwolucyjne sieci neuronowe pozwalają na przewidywanie miejsc wiązania czynników transkrypcyjnych.
- Interpretacja modelu może nakierować badaczy na nowe fakty biologiczne.
- Wyniki pochodzące z modeli typu “black-box” mogą być “generatorem hipotez”, które można potem weryfikować eksperymentalnie.