

Predykcja elementów funkcjonalnych DNA na podstawie modyfikacji histonów

Joanna Giemza

Streszczenie mojej pracy magisterskiej. Promotor: Dr Bartek Wilczynski

W ostatnich latach dokonano dużego postępu w charakteryzacji modyfikacji histonów. Modyfikacje histonów mają wpływ na wiele procesów zależnych od chromatyny, m.in. na transkrypcję. Celem mojej pracy było zbadanie, czy na podstawie danych o modyfikacjach histonów można skutecznie przewidywać lokalizację elementów funkcjonalnych DNA, takich jak egzony, introny, czy promotory. Próbowałam odróżnić je od siebie oraz od pozostałych obszarów DNA, czyli obszarów międzygenowych.

Mechanizm identyfikacji elementów funkcjonalnych DNA przez komórkę nie jest do końca wyjaśniony. Zbudowanie właściwego modelu statystycznego może pomóc zrozumieć ten proces oraz znaleźć praktyczne zastosowanie, np. jako alternatywna metoda anotacji genomu.

Problem sformułowano jako zagadnienie klasyfikacji. Klasyfikator przyporządkowywał odcinkowi DNA jedną z czterech możliwych etykiet: egzon, intron, promotor, obszar międzygenowy (kontrolny). Rozważano dwa typy klasyfikatorów: sieci bayesowskie i lasy losowe. Podejścia te, poza odmiennymi algorytmami, różniły się liczbą modyfikacji histonów branych pod uwagę podczas predykcji. Lasy losowe korzystały z pełnego, 39-elementowego zbioru modyfikacji, natomiast sieci bayesowskie ze zbioru co najwyżej 12-elementowego. Wyróżniono również najbardziej informatywne modyfikacje.

Dane pochodziły z ludzkiej linii komórkowej T CD4+. Testowano dodatkowo różne sposoby wstępnej obróbki danych, takie jak zwiększanie szerokości okna (czyli odcinka DNA poddawanego klasyfikacji) i analiza składowych głównych, redukująca wymiar danych dla klasyfikatora opartego na sieciach bayesowskich.

Okazało się, że na podstawie danych o modyfikacjach histonów można zbudować klasyfikator, który będzie istotnie lepiej niż losowo przewidywać elementy funkcjonalne DNA bez względu na aktywność genów, do których przynależą. Skoncentrowano się jednak na predykcji elementów funkcjonalnych DNA najbardziej aktywnych genów i osiągnięto jeszcze wyższą skuteczność. Najlepsze wyniki uzyskano dla klasy obszarów międzygenowych (AUC = 95%), następnie dla promotorów i egzonów (AUC = 91%), a najtrudniejszą w predykcji klasą są introny (AUC = 83%). Wyniki te pochodzą z klasyfikacji lasami losowymi dla okien o szerokości 1000 bp, która okazała się najbardziej optymalną opcją. Różnica między jakością klasyfikacji lasami losowymi a sieciami bayesowskimi jest statystycznie istotna. Sieciom bayesowskim nie pomogła redukcja wymiaru danych poprzez analizę składowych głównych. Wyniki AUC były zazwyczaj gorsze niż na danych bez preprocessingu, choć był wyjątek - klasa intronów. Najwidoczniej położenie egzonów i promotorów bardziej zależy od pojedynczych modyfikacji niż ich kombinacji liniowych wyjaśniających jak najwięcej wariacji danych.

Wyróżniono również najbardziej informatywne modyfikacje. Algorytm Boruta wskazał, że aż 37 z 39 modyfikacji histonów w zbiorze ma związek z położeniem elementów funkcjonalnych. Wyjaśnia to dużą przewagę lasów losowych nad korzystającymi z ograniczonego zbioru predyktorów sieciami bayesowskimi. Najbardziej informatywną modyfikacją według miary Z-score jest H3K36me3. Już wcześniej było wiadomo, że występuje ona przy obszarach DNA aktywnie podlegających procesowi transkrypcji. W sieciach bayesowskich znajdowała się ona wśród predyktorów egzonów, ale nie była wybierana jako predyktor intronów. Jest to zgodne z obserwacjami z pracy, że egzony są bardziej oznaczone modyfikacją H3K36me3 od intronów, i że może ona mieć zatem związek z procesem splicingu. Wśród predyktorów w sieciach bayesowskich znalazło się także 10 innych modyfikacji histonów, o zgodnych z literaturą korelacjach z elementami funkcjonalnych DNA. Sieci bayesowskie, mimo słabszej od lasów losowych jakości predykcji, okazały się pożytecznym narzędziem do charakteryzacji najbardziej informatywnych modyfikacji histonów dla poszczególnych klas.

Podobną analizę przeprowadzono na danych o embrionach muszki owocowej. Dostępne dane zawierały tylko 7 modyfikacji histonów, wzięto więc pod uwagę inne czynniki epigenetyczne takie jak białka izolatorowe, białka heterochromatynowe i deacetylazy histonów. Na tych danych również uzyskano wysoką jakość predykcji elementów funkcjonalnych DNA. Oznacza to, że możliwość przewidywania elementów funkcjonalnych nie jest charakterystyczna tylko dla linii komórkowej T CD4⁺ u człowieka. Prawdopodobnie jest to uniwersalna własność.

W przyszłości warto byłoby sprawdzić, czy klasyfikator wytrenowany na danych pewnego organizmu, jest w stanie skutecznie przewidywać elementy funkcjonalne DNA u innego organizmu. Potrzebne będą do tego dane zawierające taki sam zbiór predyktorów dla obu organizmów. Wydaje się, że istnieje szansa, by w przyszłości model do predykcji elementów funkcjonalnych DNA służył jako metoda wspomagająca anotację elementów funkcjonalnych aktywnych genów u niezbadanych jeszcze linii komórkowych i organizmów, alternatywna dla praktykowanego obecnie sekwencjonowania transkryptomów. W sieciach bayesowskich wytrenowanych na danych o embrionach muszki owocowej wśród pięciu wybieranych predyktorów znalazła się tylko jedna modyfikacja histonów (H3K36me3). Świadczy to o tym, że nie tylko modyfikacje histonów mogą być dobrymi predyktorami elementów funkcjonalnych DNA. Warto więc włączyć do zbioru danych inne czynniki epigenetyczne. Na podstawie 39 modyfikacji histonów w linii komórkowej T CD4⁺ udało się uzyskać wysoką jakość predykcji, nie dorównującą jednak skuteczności komórki. Jest zatem możliwe, że prócz modyfikacji histonów komórka wykorzystuje do identyfikacji elementów funkcjonalnych DNA inne, dodatkowe czynniki.