Autoreferat

przedstawiający opis dorobku i osiągnięć naukowych,
w szczególności określonych w art. 16 ust. 2 ustawy
o stopniach naukowych i tytule naukowym

Warszawa, 10. maja 2018

1 Imię i nazwisko

Bartosz Wilczyński

2 Posiadane dyplomy i stopnie

1. Tytuł magistra informatyki, uzyskany w 2003 roku na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego.

2. Stopień doktora nauk matematycznych w zakresie matematyki, uzyskany w 2008 roku w Instytucie Matematycznym Polskiej Akademii Nauk. Tytuł rozprawy: Stochastic Logical Networks: a mathematical framework for regulatory network reconstruction.

3 Zatrudnienie w jednostkach naukowych

1. Od 1 października 2008: adiunkt w Instytucie Informatyki na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego.

2. 2008-2011: „Postdoctoral-fellow" w European Molecular Biology Laboratory, Heidelberg, Niemcy (podczas urlopu bezpłatnego na UW).

3. X 2002- IX 2003: „research scholar" w Lawrence Livermore National Laboratory, Livermore USA.

4 Wskazanie osiągnięcia wynikającego z art. 16 ust. 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki

a) Tytuł osiągnięcia naukowego/artystycznego

Obliczeniowe metody analizy i identyfikacji niekodujących obszarów regulatorowych na skalę genomową

b) Publikacje reprezentujące osiągnięcie naukowe

[hab1] *Bartek Wilczynski, Norbert Dojer, Mateusz Patelak and Jerzy Tiuryn,* **Finding evolutionarily conserved cis-regulatory modules with a universal set of motifs**, BMC Bioinformatics, 2009, 10:82.

[hab2] *Bartek Wilczyński, Eileen Furlong*, **Dynamic CRM occupancy reflects a temporal map of developmental progression**, Molecular Systems Biology, 2010, 6:383

[hab3] *Bartek Wilczyński, Ya-Hsin Liu, Zhen Xuan Yeo, Elieen Furlong*, **Predicting Spatial and Temporal Gene Expression Using an Integrative Model of Transcription Factor Occupancy and Chromatin State**, PLoS Computational Biology, 2012, 8:12.

[hab4] *Norbert Dojer, Paweł Bednarz, Agnieszka Podsiadło, Bartek Wilczyński*, **BNFinder2: Faster Bayesian Network learning and Bayesian Classification**, Bioinformatics, 2013, 29:16 pp 2068-2070.

[hab5] *Agnieszka Podsiadło, Mariusz Wrzesień, Wiesław Paja, Witold Rudnicki, Bartek Wilczyński*, **Active enhancer positions can be accurately predicted from chromatin marks and collective sequence motif data**, 2013, BMC Systems Biology – Supplement from 24th Genome Informatics conference (GIW 2013), Singapore, 7 (Suppl 6):S16.

[hab6] *Paweł Bednarz, Bartek Wilczyński*, **Supervised learning method for predicting chromatin boundary associated insulator elements**, 2014, Journal of Bioinformatics and Computational Biology, 1442006-14

[hab7] *Julia Herman-Iżycka, Michał Własnowolski, Bartek Wilczyński*, **Taking promoters out of enhancers in sequence based predictions of tissue-specific mammalian enhancers**, 2017, BMC Medical Genomics – supplement from the 6th translational Bioinformatics Conference, Korea 2016 10(Suppl. 1):34

c) Omówienie celu naukowego/artystycznego ww. prac i osiągniętych wyników wraz z omówieniem ich ewentualnego wykorzystania

Wstęp

W ostatnich latach, od czasu odkrycia pełnych sekwencji DNA genomu ludzkiego oraz wielu innych organizmów, coraz większą uwagę przywiązuje się do analizy sekwencji niekodujących. Jednym z zaskakujących odkryć dotyczących genomu ludzkiego jest fakt, że jedynie około 1% sekwencji DNA służy kodowaniu sekwencji białkowych. Pozostałe 99% to tzw. sekwencje niekodujące, które mają inną, słabiej przez nas rozumianą, rolę w funkcjonowaniu komórek.

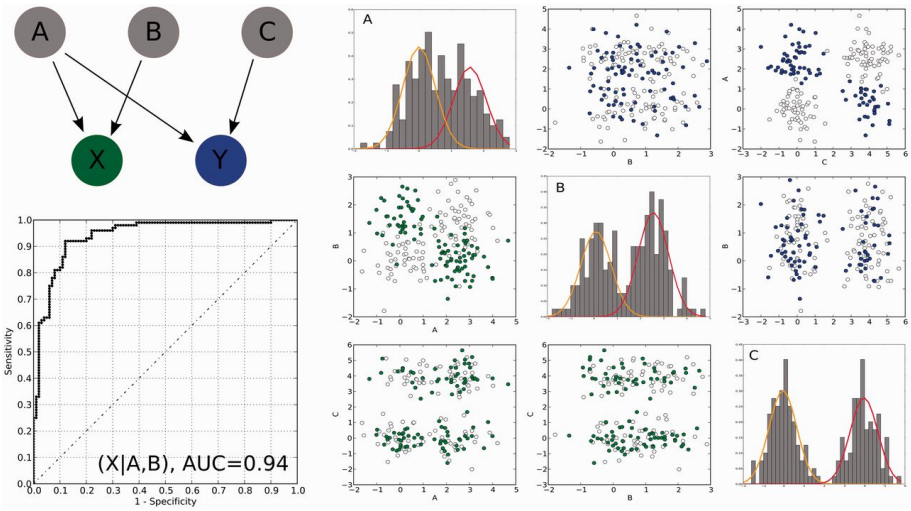
Sekwencje te są dużo mniej konserwowane w procesie ewolucji, do tego stopnia, że obecnie szacuje się, że od 20% do nawet 90% niekodującego DNA nie przenosi żadnej istotnej funkcji biologicznej[PG14]. Sytuacja ta stwarza istotny obliczeniowy problem identyfikacji funkcjonalnych sekwencji niekodujących w dużych genomach.

Mimo, że nie znamy jeszcze wszystkich funkcji biologicznych realizowanych przez sekwencje niekodujące, wiemy, że duża część z nich pełni funkcje regulatorowe, tzn. bierze udział w procesie regulacji transkrypcji poprzez wiązanie czynników transkrypcyjnych. Tego typu sekwencje regulatorowe najczęściej pełnią rolę aktywującą transkrypcję jako tzw. „enhancery" (od angielskiego wyrażenia „enhance" czyli wzmacniać). Często za regulację jednego genu odpowiada wiele takich sekwencji, które wspólnie tworzą pewnego rodzaju sieć regulatorową. Szerszy opis tych zjawisk znajduje się między innymi w artykułach przeglądowych mojego współautorstwa [P13], [p29].

Dla pełnego zrozumienia procesów związanych z regulacją ekspresji danego genu, nieodzowna jest identyfikacja jak najpełniejszego zbioru sekwencji regulatorowych z nim związanych, aby móc zidentyfikować pełen zbiór czynników transkrypcyjnych z nimi powiązanych a w konsekwencji opisać całą sieć powiązań regulacyjnych.

W związku z tym, że w typowym genomie mamy do czynienia z tysiącami genów i dziesiątkami tysięcy sekwencji regulatorowych, konieczne jest opracowanie automatycznych, obliczeniowych metod pozwalających zarówno na identyfikację samych sekwencji regulatorowych, jak i opisanie połączeń: zarówno pomiędzy czynnikami transkrypcyjnymi a obszarami regulatorowymi, przez nie wiązanymi, jak i pomiędzy obszarami regulatorowymi a genami, które są przez nie regulowane. W obu przypadkach podstawową informację stanowi dla badaczy sekwencja DNA, jednak zwykle efektywne metody obliczeniowe wykorzystują także dodatkowe informacje, często pochodzące z dodatkowych eksperymentów. W szczególności, w ostatnich latach coraz większą popularnością cieszą się metody wykorzystujące różnego rodzaju znaczniki epigenetyczne (metylację DNA, modyfikacje histonów, itp.) dla poprawy jakości wyników.

W przypadku problemu identyfikacji obszarów regulatorowych, zadanie obliczeniowe polega na opracowaniu algorytmów, które pozwalają na odróżnienie określonego zestawu sekwencji zweryfikowanych jako pozytywne przykłady funkcjonalnych sekwencji regulatorowych od zestawu kontrolnego, składającego się zwykle z sekwencji losowych, lub sztucznie wygenerowanych. Ten problem dość dobrze odpowiada standardowemu sformułowaniu problemu klasyfikacji [HT], gdzie jakość wyniku najczęściej oceniana jest w procesie walidacji krzyżowej (ang. cross-validation). Stosowane są tu często zarówno standardowe metody takie jak lasy losowe [Bre01] lub maszyny wektorów wspierających [CV95], czy też sieci Bayesowskie [P12],



Rycina 1: Wykorzystanie sieci Bayesowskich do klasyfikacji w pakiecie BNFinder. Zmienne ciągłe A, B i C opisują przynależność do zmiennych klasowych binarnych X i Y. Zmienne X i Y są opisane kolorami (odpowiednio zielony i niebieski). Z ryciny widać, że żadna ze zmiennych nie jest wystarczająca do samodzielnej klasyfikacji, ale parami pozwalają na dość dobre rozróżnienie obserwacji. Rycina zaadaptowana z [hab4]

[hab4], jak i specjalizowane metody takie jak metoda Billboard opisana w jednej z prac będących częścią dzieła habilitacyjnego [hab1].

W przypadku problemu identyfikacji połączeń pomiędzy obszarami regulatorowymi a czynnikami transkrypcyjnymi oraz genami regulowanymi, w zasadzie na obecnym etapie stosowane są głównie metody specjalizowane do tego celu. W przypadku analizy powiązań pomiędzy obszarami regulatorowymi a czynnikami transkrypcyjnymi, stosowane są zasadniczo przede wszystkim metody oparte na wykrywaniu motywów sekwencyjnych w sekwencjach DNA, lub dane z eksperymentów typu ChIPSeq. W przypadku motywów sekwencyjncych, dobre przykłady tego typu metod to np. te opisane w pracy [P18] oraz zaimplementowane w metodach Billboard [hab1], FastBill [P21], oraz module Bio.Motif pakietu Biopython [P11]. W przypadku analizy danych o wiązaniu, dobrym przykładem są prace [hab2], [hab5], [hab7], [hab6].

W kolejnych punktach postaram się skrótowo opisać jak moje badania opisane w publikacjach składających się na cykl habilitacyjny przyczyniły się do rzucenia nowego światła na niektóre aspekty wzmiankowanych tu problemów.

Nowoczesne metody klasyfikacji w oparciu o sieci Bayesowskie

W pracy [hab4] opracowaliśmy nową, ulepszoną implementację narzędzia do rekonstrukcji struktury sieci Bayesowskich pod nazwą BNFinder. Jest to narzędzie o dość szerokich zastosowaniach, natomiast w tej konkretnej wersji, będącej rozwinięciem poprzednio opublikowanej metody BNFinder [P12], jednym z głównych nowych aspektów było zaproponowanie rozwiązań pozwalających na łatwe zastosowanie sieci Bayesowskich do klasyfikacji, co ma bezpośrednie zastosowanie do problemu identyfikacji sekwencji regulatorowch. W tym przypadku zadanie jest sformułowane następująco: zakładamy, że każda cecha w problemie klasy-

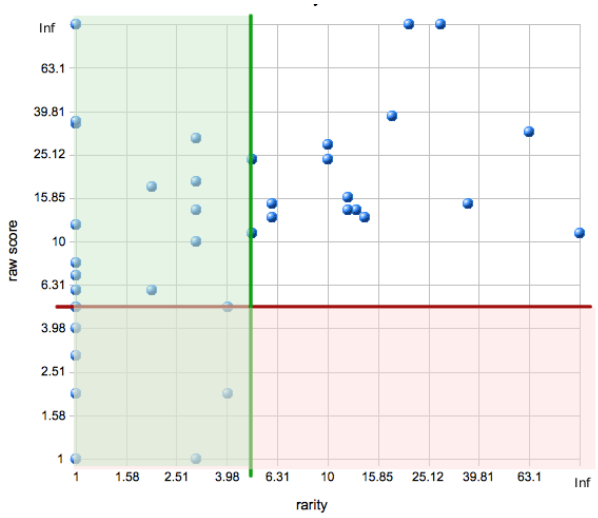
fikacji jest reprezentowana jako osobna zmienna losowa $x \in X$. Jednocześnie wszystkie zmienne klasowe są interpretowane jako zmienne losowe $y \in Y$, zaś obserwacje pochodzą z wielowymiarowego rozkładu łącznego wszystkich zmiennych $X \cup Y$. Co ważne, zakładamy, podobnie jak w standardowych sieciach Bayesowskich, że można dokonać dekompozycji rozkładu łącznego i przedstawić rozkłady brzegowe dla niektórych zmiennych w postaci rozkładów warunkowych, przy czym struktura zależności pomiędzy zmiennymi musi być grafem acyklicznym. W przypadku zastosowania sieci Bayesowskich do klasyfikacji, oczywiście interesuje nas przede wszystkim rozkład warunkowy zmiennych klasowych warunkowany cechami klasyfikowanych obiektów $P(Y|X)$, co w naturalny sposób gwarantuje acykliczność, gdyż szukana struktura grafu stanowi graf dwudzielny krawędzi pomiędzy zbiorami X i Y. Dzięki wykorzystaniu wcześniej opracowanej przez nas metody BNFinder [P12], która pozwala znaleźć optymalną strukturę takiej sieci, przy założeniu acykliczności, mogliśmy łatwo skonstruować narzędzie do klasyfikacji w oparciu o sieci Bayesowskie, także nietypowe, zawierające zmienne o rozkładach ciągłych. Dodatkowo, w pracy [hab4], zaimplementowaliśmy narzędzia pozwalające na automatyczne przeprowadzenie walidacji krzyżowej oraz wykorzystanie wielu procesorów do obliczeń równoległych. Pozwoliło to na istotne przyspieszenie działania algorytmu. Praca ta stanowiła fundament, wykorzystany zarówno w dalszych pracach należących do cyklu habilitacyjnego [hab3], [hab5], [hab6], jak i innych, prowadzonych niezależnie badaniach zarówno przeze mnie [P14], jak i innych badaczy (praca ma obecnie 10 cytowań wg bazy WebOfScience). Przykładowa sieć Bayesowska opisująca zależność dwóch różnych zmiennych klasowych (X,Y) od trzech cech (A,B,C) jest przedstawiona na Rycinie 1.

Wykrywanie obszarów regulatorowych w oparciu o motywy sekwencyjne

W pracy [hab1] zaproponowaliśmy nową metodę do wykrywania zachowanych ewolucyjnie obszarów regulatorowych. Była to pierwsza metoda pozwalająca na wykrywanie obszarów regulatorowych o luźnej strukturze motywów, opisanych w pracy [AK05]. Metoda ta opierała się na dwóch istotnych obserwacjach:

- Po pierwsze, wykorzystaliśmy funkcję oceny podobieństwa motywów występujących w dwóch homologicznych sekwencjach wykorzystującą koncepcję przesuwających się okien (ang. sliding window approach), gdzie szczegółowe położenie motywu w sekwencji nie miało znaczenia, tak długo, jak mieściło się w oknie o zadanej długości. Jedyne co liczyło się na tym etapie, to podobieństwo multizbiorów motywów występujących w sparowanych oknach.

- Po drugie, zaobserwowaliśmy, że taka prosta funkcja oceny, jest podatna na wyniki fałszywie pozytywne, nie powiązane ze specyficznymi procesami regulacji genów, a raczej wynikające z nadreprezentacji pewnych szczególnych sekwencji (np. pochodzenia transpozonowego) w genomach. Aby temu zaradzić, wprowadziliśmy nieco bardziej złożoną funkcję „rarity", która pozwalała nam ocenić, na podstawie porównań z kontrolnym zbiorem sekwencji promotorowych, na ile dany obszar stanowi rzadkość względem zbioru kontrolnego. Pozwoliło to na znaczną poprawę wyników naszej metody na danych eksperymentalnych. Przykładowe wyniki porównania prostej funkcji oceny z funkcją rarity są przedstawione na Rycinie 2.



Rycina 2: Porównanie podstawowej funkcji oceny programu billboard ("raw score") z nowo zaproponowaną funkcją "rarity", która jest empirycznym oszacowaniem częstości występowania takiego dopasowania w populacji losowych sekwencji promotorowych. Jak widać, wiele z pozytywnych sekwencji (niebieskie punkty) znajduje się w istotnym przedziale względem "rarity" (zielony obszar), a nie znajduje się w istotnym obszarze podstawowej funkcji oceny (czerwony obszar).

Prace nad tą metodą były później jeszcze kontynuowane w pracach [DBT11] oraz [P21].

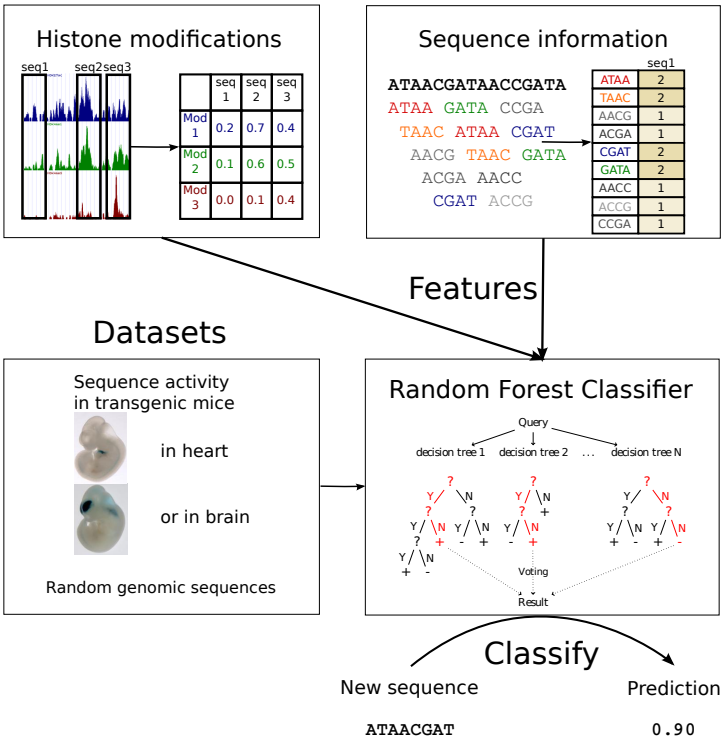
Wykrywanie obszarów regulatorowych w oparciu o eksperymentalne dane wysokoprzepustowe

Oprócz danych dotyczących motywów w sekwencjach DNA, często wykorzystywane są dane o wiązaniu czynników transkrypcyjnych do DNA. Najczęściej są to dane pochodzące z eksprymentów typu immunostrącania chromatyny (ang. Chromatin Immuno Precipitation, w skrócie ChIP), przeprowadzane obecnie w postaci wysokoprzepustowej przy pomocy mikromacierzy (ChIP-on-Chip [LRR+02]) lub sekwencjonowania nowej generacji (Chip-Seq [BCC+07]). Dzięki tym metodom, możemy zidentyfikować zarówno miejsca wiązania czynników transkrypcyjnych, jak i przy pomocy modyfikacji histonów, opisać epigenetyczny kontekst obszarów regulatorowych, co pozwala na dość dobry opis całogenomowego katalogu aktywnych obszarów regulatorowych w zadanych warunkach przy pomocy kilku eksperymentów ChIP.

Niestety, jak większość metod biochemicznych, metody typu ChIP, są pomiarami zawierającymi stosonkowo niewielki stosunek sygnału do szumu, co powoduje, że nawet posiadając kilka powtórzeń eksperymentu, nie jest sprawą oczywistą dokładna identyfikacja obszarów regulatorowych od innych sekwencji niekodujących.

W związku z tym, w typowym podejściu obliczeniowym do tego problemu, stosuje się najczęściej klasyfikatory oparte na połączeniu cech sekwencyjncych i pomiarów ChIP oraz dane o cechach sekwencyjnych w postaci szczegółowych motywów DNA, lub prostych zliczeń podsekwencji długości k. Oczywiście, klasyfikacja przebiega w procesie uczenia pod nadzorem,

na podstawie zbioru uczącego opartego na wiedzy eksperckiej pochodzącej od biologów. Dobrze tego typu podejścia ilustruje Rycina 3.



Rycina 3: Schemat metody wykorzystania metod klasyfikacji sekwencji DNA zaadaptowany z pracy [hab7]. Wykorzystując modyfikacje histonów (lewy górny róg) i motywy sekwencyjne (prawy górny róg) jako cechy, zaś opisane wyniki eksperymentów typu "enhancer-reporter" (lewy dolny róg) jako zmienną klasową, konstruujemy klasyfikator (prawy dolny róg), w tym przypadku korzystający z metody lasów losowych, który możemy użyć do klasyfikowania nowych sekwencji.

W moim dorobku znajduje się kilka prac wykorzystujących podobny schemat klasyfikacji do wykrywania funkcjonalnych sekwencji regulatorowych w niekodujących częściach genomu.

Między innymi, w pracy [hab5], napisana wspólnie z Agnieszką Podsiadło, Mariuszem Wrześniem, Wiesławem Pają i Witoldem Rudnickim, wykorzystuje metody lasów losowych, sieci Bayesowskie i maszyny wektorów wspierających do predykcji enhancerów w genomie *D. melanogaster*. Nasza metoda wykorzystania jednocześnie danych sekwencyjnych i modyfikacji histonów okazała się dawać wyraźnie lepsze efekty niż wcześniejsza praca [P14], w której wykorzystywaliśmy sieci Bayesowskie do podobnego problemu, jednak bez użycia motywów sekwencyjnych. Wzrost jakości klasyfikacji, mierzonej jako pole powierzchni pod krzywą ROC w walidacji krzyżowej był istotny - z 0.80 do 0.97. Dodatkowo, dzięki wykorzystaniu pakietu Boruta [KR+10], byliśmy w stanie zinterpretować relatywny wkład różnych cecch w działanie klasyfikatora.

W pracy [hab6], wspólnie z Pawłem Bednarzem wykorzystaliśmy podobną metodologię, jednak w zupełnie innym zadaniu. Mianowicie opracowaliśmy klasyfikator, pozwalający na

rozpoznanie tzw. sekwencji izolatorowych, tzn. funkcjonujących jako granice blokujące interakcje enhancerów z genami. Jest to istotne zadanie, gdyż bez identyfikacji izolatorów, trudno jest zbudować modele łączące enhancery z genami. Metodologicznie praca była podobna do poprzedniej, wykorzystywała przede wszystkim lasy losowe jako narzędzie klasyfikacji i pakiet Boruta do określania relatywnej ważności cech, jednak w związku z tym, że dane były innej natury (krótkie słowa długości k zamiast motywów miejsc wiązania czynników transkrypcyjnych i dane ChIP dla białek izolatorowych zamiast modyfikacji histonów) wymagała dość dużego wysiłku włożonego w dostrojenie parametrów metody. Była to wg naszej wiedzy pierwsza praca stosująca tego typu podejście do obszarów izolatorowych, później nasze podejście zostało zastosowane także do podobnych danych w innych gatunkach [MAS15].

Także w pracy [hab7], wykorzystaliśmy klasyfikatory oparte na lasach losowych, jednak tutaj podejście było nieco bardziej złożone, gdyż naszym głównym celem nie była klasyfikacja obszarów regulatorowych względem sekwencji kontrolnych, ale rozróżnienie sekwencji regulatorowych różnego rodzaju, w szczególności pomiędzy enhancerami aktywnymi w różnych tkankach oraz promotorami genów. Głównym wynikiem tej pracy z punktu widzenia biologicznego, było wykazanie, że dla dokładnego rozróżnienia pomiędzy różnymi grupami enhancerów jest istotne rozważenie cech sekwencyjnych występujących także w promotorach. Ostatecznie najlepsze wyniki dawała metoda łącząca dwa klasyfikatory uczone na różnych zbiorach uczących: enhancerach i promotorach. Połączenie tych dwóch klasyfikatorów było podejściem nie stosowanym wcześniej w tym kontekście i dającym istotnie lepsze wyniki niż wcześniejsze metody.

Analiza funkcjonalna obszarów regulatorowych w oparciu o dane wysokoprzepustowe

Identyfikacja obszarów regulatorowych jest warunkiem koniecznym do budowy złożonych modeli regulacji genów, ale nie wystarczającym. Konieczne jest także identyfikacja powiązań czynników transkrypcyjnych z obszarami regulatorwymi oraz opis ich funkcji. W pracy [hab2] opublikowanej wspólnie z Eileen Furlong, zajęliśmy się właśnie analizą relacji pomiędzy sygnałem sekwencyjnym opisanym przy pomocy motywów miejsc wiązania czynników transkrypcyjnych a sygnałem z eksperymentów ChIP dla tych samych czynników. Była to analiza szczególnego zbioru danych, bo zawierającego zmieniające się w czasie pomiary wiązania, ówcześnie zupełnie unikalne. Fakt, że próbowaliśmy powiązać zmieniające się w czasie wiązanie czynników do sekwencji, przy jednocześnie niezmiennych motywach DNA wymagało nowatorskiej analizy danych. Wykorzystaliśmy tu metodę TRAP [RKMV06] do pomiaru ilościowego powinowactwa zadanej sekwencji do motywu oraz porównawcze analizy pomiędzy grupami sekwencji wykazującymi zróżnicowane zachowanie jeśli chodzi o wiązanie czynników transkrypcyjnych w czasie, aby udowodnić, że istotnymi elementami dla funkcji enhancerów jest zarówno kombinatoryczne wiązanie ko-czynników transkrypcyjnych jak i kontekst epigenetyczny. Prace te zaowocowały nie tylko licznymi odwołaniami w literaturze (praca ma obecnie 24 cytowania wg bazy Web Of Science) ale także powstaniem narzędzi do analizy funkcjonalnej grup genów [KHIW16].

Złożone modele predykcji w oparciu o klasyfikatory

Swego rodzaju zwieńczeniem prac wspomnianych wcześniej jest praca [hab3], która opisuje złożony model, pozwalający na dokładną (co do czasu i miejsca) predykcję ekspresji genów w rozwijającym się zarodku. Połączyliśmy w niej dane o pozycjach enhancerów, powiązaniach czynników transkrypcyjnych, izolatorach DNA i modyfikacjach histonów w spójny model probabilistyczny, który pozwolił nam przewidzieć ekspresję setek genów w różnych tkankach podczas procesu różnicowania się mięśni w rozwoju zarodkowym *D. melanogaster*. Była to w naturalny sposób praca zespołowa, wymagająca najpierw zebrania danych do procesu uczenia, później konstrukcji modelu, dobrania parametrów, przeprowadzenia procesu uczenia i weryfikacji eksperymentalnej predykcji.

Model oparty był na dwóch podzespołach (przedstawionych także na Rycinie 4.):

- Klasyfikatorze Bayesowskim, który na podstawie pomiarów siły wiązania czynników transkrypcyjnych do sekwencji DNA pozwalał określić prawdopodobieństwo aktywności danej sekwencji w różnych tkankach,
- Prostym modelu probabilistycznym, który wiązał aktywację genów z niezależnymi zdarzeniami aktywacji enhancerów, położeniem izolatorów oraz stanem epigenetycznym promotora.

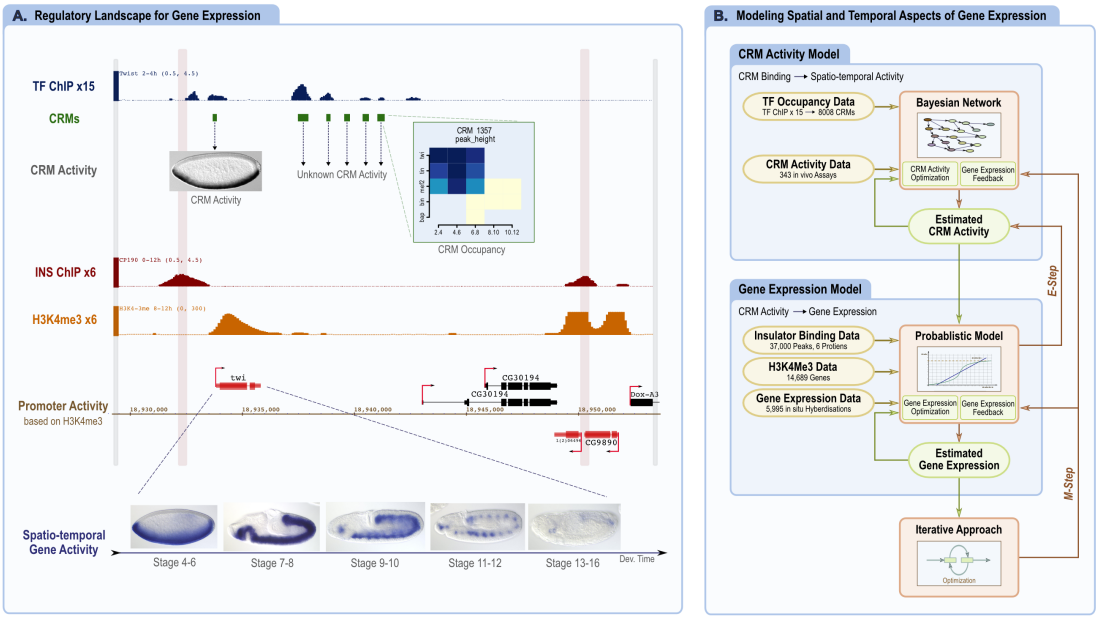
Dzięki takiej konstrukcji modelu, byliśmy w stanie opisać funkcję wiarygodności (Log-likelihood) i przeprowadzić procedurę maksymalizacji wartości oczekiwanej (Expectation Maximization) prowadzącą do znalezienia modelu lokalnie optymalnego. Dzięki temu, że procedura optymalizacji została wykonana wielokrotnie, upewniliśmy się, że znalezione optimum lokalne jest atraktorem dla dużego zakresu wartości początkowych.

Przy pomocy tego modelu osiągnęliśmy nie tylko obiektywnie wysoką skuteczność predykcji zweryfikowaną eksperymentalnie na niezależnym zbiorze testowych genów, ale także dokonaliśmy pewnych przewidywań natury biologicznej, zweryfikowanych później w kolejnych pracach. W szczególności, nasza praca wskazała na konieczność częstej regulacji wielu genów przez ten sam enhancer oraz konieczność istnienia oddziaływań dalekiego zasięgu pomiędzy genami a enhancerami dla prawidłowej regulacji genów w rozwoju zarodkowym.

5 Omówienie pozostałych osiągnięć naukowo - badawczych

Moja praca nad doktoratem dotyczyła przede wszystkim matematycznego modelowania sieci regulatorowych i wyniki te oprócz rozprawy doktorskiej były publikowane głównie w wydawnictwach konferencyjnych ([p26], [p27]). Już w trakcie prac nad doktoratem interesowałem się także bardziej praktycznymi problemami identyfikacji grup czynników transkrypcyjnych związanych z profilami ekspresji genów ([p25], [P8], [P10]), a także ogólnym problemem identyfikacji struktury sieci Bayesowskich - zarówno w zastosowaniu Dynamicznych sieci Bayesowskich do sieci regulacji genów ([P9]) jak i w ogólnym przypadku sieci acyklicznych [P12].

Po uzyskaniu stopnia doktora w 2008 roku podjąłem badania nad szeroko pojętymi obliczeniowymi metodami w zastosowaniu do wysoko-przepustowych danych biologicznych.



Rycina 4: Schemat złożonego modelu probabilistycznego przewidującego ekspresję genów w oparciu o wiązanie czynników transkrypcyjnych, modyfikacje histonów, przewidywaną aktywność obszarów regulatorowych i ich położenie względem promotorów i izolatorów. Rycina zaadoptowana z pracy [hab3]

Głównym nurtem moich zastosowań pozostał oczywiście problem regulacji genów, jednakże pod względem metod moja działaność skupiała się na kilku różnych podejściach do tego tematu:

- Dalszy rozwój metod związanych z wyszukiwaniem obszarów regulatorowych, w szczególności moduł Bio.Motif biblioteki Biopython [P11], rozwój równoległych algorytmów do rekonstrukcji sieci Bayesowskich [p31], zaawansowanych metod określania optymalnych progów wykrycia miejsc wiązania czynników transkrypcyjnych dla popularnych baz danych [P18]oraz dalszy rozwój narzędzi do wykrywania obszarów regulatorowych zachowanych ewolucyjnie w postaci narzędzia FastBill [P21].

- Rozwój zaawansowanej metody do wykrywania mutacji w danych ChIP-Seq i wspomaganej rekonstrukcji sekwencji genomu RECORD (wspólnie z drem Norbertem Dojerem i drem Krisztianem Buzą) [P19]

- Zastosowania metod opracowanych w pracach będących częścią dzieła habilitacyjnego w projektach prowadzonych w ścisłej współpracy z biologami pracującymi nad regulacją genów u *D. melanogaster* : dr Eileen Furlong:[P13], [P14], [P15]oraz prof. Davidem Arnostim [P17]

- Zastosowania zaawansowanych metod analizy danych wysokoprzepustowych z sekwencjonowania DNA we współpracy z biologami zajmującymi się regulacją genów w *A. thaliana*: [P16], [P20]

- Zastosowania metod statystycznej analizy sekwencji DNA do wykrywania istotnych mutacji u pacjentów z nowotworami we współpracy z grupą dra Tomasza Wilanowskiego: [p30], [P22], [P24]

- Wspólnie z grupą Minny Kaikkonen z Kuopio analizowaliśmy rolę struktury chromatyny w regulacji genów, co zaowocowało opracowaniem nowych metod analizy macierzy kontaktów chromosomowych oraz publikacją w czasopiśmie Nucleic Acids Research[P23]

Podsumowując, moja działalność naukowa poza publikacjami z głównego cyklu składającego się na osiągnięcie habilitacyjne, była bardzo różnorodna i obfitowała w różnorakie współprace z grupami biologów eksperymentalnych. Moim zdaniem, te „poboczne" współprace miały bardzo istotny wpływ także na główny nurt moich badań oraz na obierane przeze mnie kierunki rozwoju narzędzi obliczeniowych. Tego rodzaju bliskie współprace z biologami, jakkolwiek absorbujące czasowo, pozwalają lepiej zrozumieć naturę problemów biologicznych i dostosować do niej kształt ostatecznie wybieranych rozwiązań informatycznych.

Poniżej, w części 5.1, znajduje się lista wszystkich moich publikacji z podziałem na te opublikowane w czasopismach indeksowanych w JCR (oznaczone literą P i kolejnymi numerami) i pozostałych (oznaczonych literą p i kolejnym numerem), nie licząc tych, które zostały wymienione w punkcie 4 jako części składowe dzieła habilitacyjnego. Wiele z tych prac, pobocznych z punktu widzenia habilitacji, jest ważnych dla szeroko rozumianego środowiska naukowego. Największy wpływ, z punktu widzenia odzewu środowiska w postaci licznych cytowań, wywarły następujące publikacje:

[P11] Narzędzie biopython stosowane masowo przez bioinformatyków z całego świata, obecnie ma 704 cytowania według bazy Web of Science,

[P14] Predykcja enhancerów przy pomocy sieci Bayesowskich (bnfinder) na podstawie danych BITS-CHIP opublikowana w czasopiśmie Nature Genetics, wg bazy Web of Science była dotąd cytowana 215 razy,

[P9] Nasza metoda rekonstrukcji sieci regulatorowych przy pomocy dynamicznych sieci Bayesowskich była cytowana 90-krotnie (wg WoS).

5.1 Pozostałe publikacje naukowe

5.1.1 W czasopismach indeksowanych w JCR

Przed uzyskaniem stopnia doktora:

[P8] *Hvidsten, Torgeir Roden; Wilczynski, Bartosz; Kryshtafovych, Andriy; Tiuryn, Jerzy; Komorowski, Jan; Fidelis, Krzysztof*, **Discovering regulatory binding-site modules using rule-based learning**, Genome Research, 2005, 15(6) pp. 856-866.

[P9] *Dojer, Norbert; Gambin, Anna; Mizera, Andrzej; Wilczynski, Bartek; Tiuryn, Jerzy*, **Applying dynamic Bayesian networks to perturbed gene expression data**, 2006, BMC Bioinformatics, 7:249

[P10] *Wilczynski, Bartek; Hvidsten, Torgeir R.; Kryshtafovych, Andriy; Tiuryn, Jerzy; Komorowski, Jan; Fidelis, Krzysztof*, **Using local gene expression similarities to discover regulatory binding site modules**, 2006, BMC Bioinformatics, 7:505

Po uzyskaniu stopnia doktora:

[P11] *Cock, Peter J. A.; Antao, Tiago; Chang, Jeffrey T.; Chapman, Brad A.; Cox, Cymon J.; Dalke, Andrew; Friedberg, Iddo; Hamelryck, Thomas; Kauff, Frank; Wilczynski, Bartek; de Hoon, Michiel J. L.*, **Biopython: freely available Python tools for computational molecular biology and bioinformatics**, 2009, Bioinformatics, 25(11) pp. 1422-1423

[P12] *Wilczynski, Bartek; Dojer, Norbert*, **BNFinder: exact and efficient method for learning Bayesian networks**, 2009, Bioinformatics, 25(2), pp. 286-287,

[P13] *Wilczynski, Bartek; Furlong, Eileen E. M*, **Challenges for modeling global gene regulatory networks during development: Insights from Drosophila**, 2010, Developmental Biology, 340(2) pp. 161-169.

[P14] *Bonn, Stefan; Zinzen, Robert P.; Girardot, Charles; Gustafson, E. Hilary; Perez-Gonzalez, Alexis; Delhomme, Nicolas; Ghavi-Helm, Yad; Wilczynski, Bartek; Riddell, Andrew; Furlong, Eileen E. M.*, **Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development**, 2012, Nature Genetics, 44(2) pp. 148-156

[P15] *Ciglar, Lucia; Girardot, Charles; Wilczynski, Bartek; Braun, Martina; Furlong, Eileen E. M.*, **Coordinated repression and activation of two transcriptional programs stabilizes cell fate during myogenesis**, Development, 2014, 141(13), pp. 2633-2643.

[P16] *Rutowicz, Kinga; Puzio, Marcin; Halibart-Puzio, Joanna; Lirski, Maciej; Kotlinski, Maciej; Kroten, Magdalena A.; Knizewski, Lukasz; Lange, Bartosz; Muszewska, Anna; Sniegowska-Swierk, Katarzyna; Koscielniak, Janusz; Iwanicka-Nowicka, Roksana; Buza, Krisztian; Janowiak, Franciszek; Zmuda, Katarzyna; Joesaar, Indrek; Laskowska-Kaszub, Katarzyna; Fogtman, Anna; Kollist, Hannes; Zielenkiewicz, Piotr; Tiuryn, Jerzy; Siedlecki, Pawel; Swiezewski, Szymon; Ginalski, Krzysztof; Koblowska, Marta; Archacki, Rafal; Wilczynski, Bartek; Rapacz, Marcin; Jerzmanowski, Andrzej*, **A Specialized Histone H1 Variant Is Required for Adaptive Responses to Complex Abiotic Stress and Related DNA Methylation in Arabidopsis**, Plant Physiology, 2015, 169(3), pp. 2080-2101

[P17] *Wei, Yiliang; Mondal, Shamba Sankar; Mouawad, Rima; Wilczynski, Bartek; Henry, R. William; Arnosti, David N.*, **Genome-Wide Analysis of Drosophila RBf2 Protein Highlights the Diversity of RB Family Targets and Possible Role in Regulation of Ribosome Biosynthesis**, G3-Genes, Genomes, Genetics, 2015, 5(7) pp. 1503-1515

[P18] *Dabrowski, Michal; Dojer, Norbert; Krystkowiak, Izabella; Kaminska, Bozena; Wilczynski, Bartek*, **Optimally choosing PWM motif databases and sequence scanning approaches based on ChIP- seq data**, BMC Bioinformatics, 2015, 16:140.

[P19] *Buza, Krisztian; Wilczynski, Bartek; Dojer, Norbert*, **RECORD: Reference-Assisted Genome Assembly for Closely Related Genomes**, International Journal of Genomics, 2015, 563482

[P20] *Archacki, Rafal; Yatusevich, Ruslan; Buszewicz, Daniel; Krzyczmonik, Katarzyna; Patryn, Jacek; Iwanicka-Nowicka, Roksana; Biecek, Przemyslaw; Wilczynski, Bartek; Koblowska, Marta; Jerzmanowski, Andrzej; Swiezewski, Szymon*, **Arabidopsis SWI/SNF chromatin remodeling complex binds both promoters and terminators to regulate gene expression**, Nucleic Acids Research, 2017, 45(6), pp. 3116-3129

[P21] *Wilczynski, Bartek; Tiuryn, Jerzy*, **FastBill: An Improved Tool for Prediction of Cis-Regulatory Modules**, Journal of Computational Biology, 2017, 24(3),pp 193-199.

[P22] *Pawlak, Magdalena; Kikulska, Agnieszka; Wrzesinski, Tomasz; Rausch, Tobias; Kwias, Zbigniew; Wilczynski, Bartek; Benes, Vladimir; Wesoly, Joanna; Wilanowski, Tomasz*, **Potential protective role of Grainyhead-like genes in the development of clear cell renal cell carcinoma**, Molecular Carcinogenesis, 2017, 56(11) pp.2414-2423

[P23] *Niskanen, Henri; Tuszynska, Irina; Zaborowski, Rafal; Heinäniemi, Merja; Ylä-Herttuala, Seppo; Wilczynski, Bartek; Kaikkonen, Minna U*, **Endothelial cell differentiation is encompassed by changes in long range interactions between inactive chromatin regions**, Nucleic acids research, 2017,online publication gkx1214

[P24] *Kikulska, Agnieszka; Rausch, Tobias; Krzywinska, Ewa; Pawlak, Magdalena; Wilczynski, Bartek; Benes, Vladimir; Rutkowski, Piotr; Wilanowski, Tomasz*, **Coordinated expression and genetic polymorphisms in Grainyhead-like genes in human non-melanoma skin cancers**, BMC Cancer, 2018, 18(1):23

5.1.2 W wydawnictwach spoza listy JCR

Przed uzyskaniem stopnia doktora:

[p25] *Wilczynski, B; Hvidsten, T; Kryshtafovych, A; Stubbs, L; Komorowski, J; Fidelis, K*, **A rule-based framework for gene regulation pathways discovery**, PROCEEDINGS OF THE 2003 IEEE BIOINFORMATICS CONFERENCE, 2003, Stanford, pp 435-436.

[p26] *Wilczynski, Bartek; Tiuryn, Jerzy*, **Regulatory network reconstruction using stochastic logical networks**, PROCEEDINGS of the conference on COMPUTATIONAL METHODS IN SYSTEMS BIOLOGY, 2006, Trento, Lecture notes in bioinformatics, 4210, pp. 142-154.

[p27] *Wilczynski, Bartek; Tiuryn, Jerzy*, **Reconstruction of mammalian cell cycle regulatory network from Microarray data using Stochastic logical networks**, 2007, PROCEEDINGS of the conference on COMPUTATIONAL METHODS IN SYSTEMS BIOLOGY, Edinburgh, Lecture notes in bioinformatics, 4695, p. 121-135

[p28] *Wilczyński, Bartek*, **A stochastic extension of R. Thomas regulatory network modelling**,Banach Center Publications, 2008, Vol. 80,pp. 271–276.

Po uzyskaniu stopnia doktora:

[p29] *Wilczyński, Bartek and Hvidsten, Torgeir R*, **A Computer Scientist's Guide to the Regulatory Genome**, Fundamenta Informaticae, 2010, Vol.103 (1), pp. 323–332.

[p30] *Kikulska, Agnieszka; Wilczynski, Bartosz.; Rausch, Tobias; Benes, Vladimir; Rutkowski, Piotr; Wilanowski, Tomasz*, **Reduced expression of GRHL genes in human non-melanoma skin cancers**, European Journal of Cancer, 2014, 50 pp S34-35

[p31] *Frolova, Alina; Wilczynski, Bartek*, **Fast Parallel Bayesian Networks Reconstruction with BNFinder**, 2014, PROCEEDINGS IWBBIO 2014: INTERNATIONAL WORK-CONFERENCE ON BIOINFORMATICS AND BIOMEDICAL ENGINEERING, 1-2 pp. 1179-1184.

Pozostała literatura

[AK05] David N Arnosti and Meghana M Kulkarni. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of cellular biochemistry*, 94(5):890–898, 2005.

[BCC+07] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[DBT11] Norbert Dojer, Przemyslaw Biecek, and Jerzy Tiuryn. Bi-billboard: symmetrization and careful choice of informant species results in higher accuracy of regulatory element prediction. *Journal of Computational Biology*, 18(6):809–819, 2011.

[HT] Tibshirani Hastie and R Tibshirani. & friedman, j.(2008). the elements of statistical learning; data mining, inference and prediction.

[KHIW16] Kamil Koziara, Julia Herman-Izycka, and Bartek Wilczynski. Bio.ontology - python tools for enrichment analysis and visualization of ontologies. *bioRxiv*, 2016.

[KR+10] Miron B Kursa, Witold R Rudnicki, et al. Feature selection with the boruta package. *J Stat Softw*, 36(11):1–13, 2010.

[LRR+02] Tong Ihn Lee, Nicola J Rinaldi, François Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, et al. Transcriptional regulatory networks in saccharomyces cerevisiae. *science*, 298(5594):799–804, 2002.

[MAS15] Benjamin L Moore, Stuart Aitken, and Colin A Semple. Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. *Genome biology*, 16(1):110, 2015.

[PG14] Alexander F Palazzo and T Ryan Gregory. The case for junk dna. *PLoS genetics*, 10(5):e1004351, 2014.

[RKMV06] Helge G Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. Predicting transcription factor affinities to dna from a biophysical model. *Bioinformatics*, 23(2):134–141, 2006.