

MICHAŁ DĄBROWSKI, BOŻENA KAMIŃSKA

*Pracownia Regulacji Transkrypcji
Instytut Biologii Doświadczalnej PAN
Pasteura 3, 02-093 Warszawa
E-mail: m.dabrowski@nencki.gov.pl*

OD WZORÓW EKSPRESJI GENÓW I MOTYWÓW REGULATOROWYCH DO PREDYKCJI I MODELOWANIA EKSPRESJI GENÓW – WYZWANIA I NADZIEJE

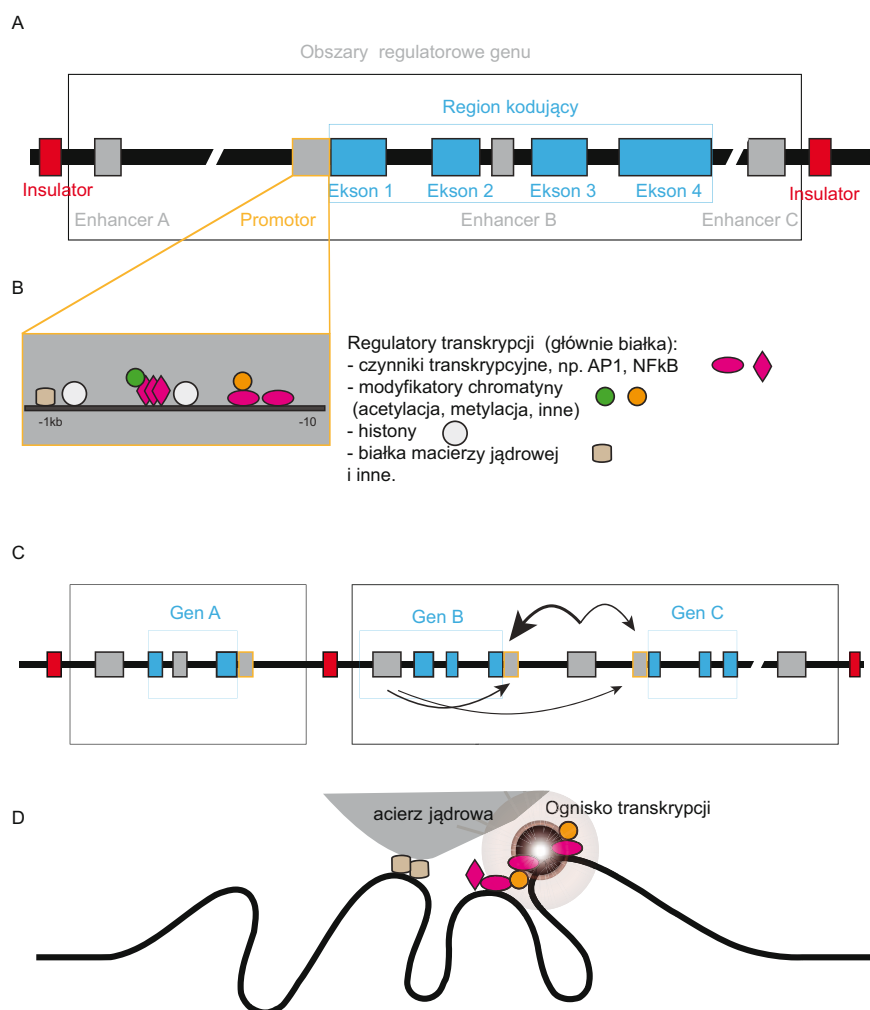
WPROWADZENIE

Zróźnicowanie ekspresji genów, pomiędzy komórkami zawierającymi identyczny materiał genetyczny i będącymi produktami podziału tej samej komórki, stanowi podstawę procesów różnicowania komórek w rozwoju organizmu oraz determinuje specyficzny fenotyp i funkcje komórek. Nowoczesne, wielkoskalowe metody analizy ekspresji genów, rozwinięte w ostatnich latach, pozwoliły po raz pierwszy na porównanie różnych rodzajów tkanek w stanach fizjologicznych i patologicznych na poziomie całych transkryptomów. Cechą charakterystyczną zmian ekspresji genów w rozwoju, w trakcie procesów fizjologicznych i patologicznych jest równoczesna współregulacja setek, a być może nawet tysięcy genów. Również odpowiedź komórek na specyficzny bodziec (np. uszkodzenie DNA, inicjacja śmierci komórkowej, działanie cytokin lub czynników wzrostowych) zwykle obejmuje zmiany setek/tysięcy genów. Poznanie mechanizmów współregulacji ekspresji wielu genów stanowi jeden z najbardziej fascynujących obszarów badań biologicznych. Postęp w tej dziedzinie wydaje się kluczowy dla lepszego zrozumienia procesów rozwoju (w tym regeneracji) oraz innych długotrwałych zmian czynnościowych naszych organizmów, takich jak uczenie się, choroby przewlekłe, proces starzenia się.

Klasyczne metody doświadczalne (opracowane w celu badania pojedynczych genów) pozwoliły na szczegółowe zbadanie regulacji ekspresji setek genów, w różnych orga-

nizmach modelowych i sytuacjach doświadczalnych. Wyłaniający się z tych badań obraz, choć niezwykle złożony, pozwala na pewne uogólnienia. Regulacja ekspresji niemal każdego genu jest złożonym wielopoziomowym procesem, w którym regulacji podlega ilość produktu danego genu zarówno na poziomie szybkości tworzenia jak i degradacji jego transkryptu (mRNA). O poziomie produkcji kodowanego przez gen białka decyduje wypadkowa między syntezą białka a jego degradacją. Stabilność białka oraz jego aktywność podlegają w komórkach dodatkowej regulacji poprzez modyfikacje potranslacyjne.

W większości procesów biologicznych o poziomie transkrypcji genu decyduje inicjacja transkrypcji. Przyjmuje się, że regiony regulatorowe ulokowane są w obszarach 10 000–50 000 par zasad od miejsca startu transkrypcji i wyróżnia się w nich sekwencje wzmacniające (ang. enhancers) lub wyciszające ekspresję (ang. silencers), promotor regulatorowy i rdzeń promotora (ang. core promoter), zwykle znajdujący się w sąsiedztwie miejsca startu transkrypcji (Ryc. 1). W obrębie obszarów regulatorowych znajduje się wiele miejsc wiążących czynniki transkrypcyjne, które rozpoznają określone sekwencje DNA, wiążą się z nimi i kierują podstawowy aparat transkrypcyjny do najbliższego promotora inicjując proces transkrypcji mRNA. Inicjacja syntezy mRNA w komórkach eukariotycznych jest procesem wieloetapowym i wymaga skoordynowanego współdziała-



Ryc. 1. Organizacja obszarów regulatorowych organizmów tkankowych.

A. Obszary cis-regulatorowe genu: promotor i obszary wzmacniające ekspresję (enhancery). B. Wspólną cechą promotorów i obszarów wzmacniających (także wyciszających) jest nagromadzenie miejsc wiążących czynniki transkrypcyjne. C. Oddziaływania pomiędzy promotorami, enhancerami i insulatorami. Te ostatnie izolują geny od wpływu dystalnie położonych obszarów cis-regulatorowych. D. Oddziałujące ze sobą sekwencje regulatorowe są związane z białkami. Oddziaływania te zachodzą w trójwymiarowej strukturze jądra, którą aktywnie kształtują.

nia podstawowych czynników transkrypcyjnych i RNA polimerazy II (RNAPII). Rdzeń promotora jest definiowany jako minimalny fragment DNA zdolny do inicjowania transkrypcji przez RNAPII na niskim poziomie w nieobecności aktywatorów. Rdzeń promotora to typowo obszar około 40 nukleotydowy poniżej i powyżej miejsca startu transkrypcji, który wykazuje duże zróżnicowanie sekwencji. Stwierdzono, że sekwencje TATA box oraz INR (ang. initiator) są w stanie kierować transkrypcję przez RNAPII w nieobec-

ności innych elementów regulatorowych (ORPHANIDES i REINBERG 2002, SMALE i KADONAGA 2003). Analiza bioinformatyczna genomów wielokomórkowych Eukaryota wskazuje, że obecność sekwencji zwanej TATA box w promotorach genów była przeceniana w przeszłości i obecnie przyjmuje się, że np. większość ludzkich genów jest pozbawiona sekwencji TATA box (GROSS i OELGESCHLAGER 2006). Oprócz rdzenia promotora wyróżnia się promotor regulatorowy, który zawiera miejsca wiązania aktywatorów transkrypcji.

Aktywatory (zwykle indukowane czynniki transkrypcyjne) zwiększają efektywność transkrypcji w dwojaki sposób: po pierwsze poprzez interakcje białko-białko zwiększają wiązanie się maszyny transkrypcyjnej z DNA w promotorze; po drugie niektóre aktywatory stabilizują konformację kompleksu TFIIA-TFIID-DNA w promotorze.

Regulatorowe czynniki transkrypcyjne (TFs) zbudowane są zwykle z domeny wiążącej specyficznie docelową sekwencję DNA oraz domeny transkrypcyjnej, która aktywuje lub hamuje transkrypcję. W zależności od budowy domen wiążących DNA czynniki transkrypcyjne pogrupowano w rodziny, których zidentyfikowano około 100 (KUMMERFELD i TEICHMANN 2006). Różnorodność regulatorowych czynników transkrypcyjnych odzwierciedla stopień skomplikowania organizmów: np. genom jednokomórkowych drożdży *Saccharomyces cerevisiae* koduje około 200 przewidywanych TFs (HARBISON i współaut. 2004), co stanowi 3% wszystkich genów kodujących białka. Względnie prosty organizm wielokomórkowy taki jak nicienie *Caenorhabditis elegans* ma 934 genów kodujących przewidywane TFs, co stanowi 5% wszystkich genów kodujących białka, a u wyższych organizmów np. u człowieka geny kodujące TFs stanowią 10% (LEVINE i TJIAN 2003).

Postuluje się, że miejsca wiązania regulatorowych czynników transkrypcyjnych występują często w skupieniach tworząc regulatorowe moduły (DAVIDSON 2006). Hipoteza modułowości obszarów regulatorowych została potwierdzona z użyciem metod bioinformatycznych i wykorzystana przez liczne grupy do znalezienia obszarów *cis*-regulatorowych tworzących promotory regulatorowe i obszary wzmacniające lub wyciszające ekspresję (AERTS i współaut. 2003, SHARAN i współaut. 2003, GUPTA i LIU 2005, HALLIKAS i współaut. 2006). Obszary regulatorowe promotora mają budowę modułową, tzn. zawierają zwykle miejsca wiązania licznych czynników transkrypcyjnych, które ściśle współpracują ze sobą (SEGAL i współaut. 2008). Przewiduje się, że do 10% genomu koduje czynniki transkrypcyjne, które wiążąc się w różnych kombinacjach w promotorach genów determinują inicjację oraz określają intensywność transkrypcji. Często te same czynniki transkrypcyjne wiążą się do promotorów różnych genów. Modułowy charakter obszarów regulatorowych genu jest cechą bardzo ważną, gdyż oznacza, że często wyłączenie pojedynczego czynnika transkrypcyjnego sprawia, że cały moduł transkrypcyjny staje się niefunkcjonalny.

CZY PODOBNY PROFIL TKANKOWY LUB JEDNOCZESNA EKSPRESJA WIELU GENÓW ODZWIERCIEDLA PROCESY ICH WSPÓŁREGULACJI?

Badanie zagadnień współregulacji wymaga równoczesnych pomiarów ekspresji wielu genów. Rozpowszechnienie się w ciągu ostatnich kilkunastu lat metod równoczesnego pomiaru ekspresji mRNA tysięcy genów – tzw. profilowania ekspresji, głównie w oparciu o technologie mikromacierzowe, dostarczyło danych koniecznych do badań współregulacji na poziomie mRNA (BARRETT i współaut. 2009, PARKINSON i współaut. 2009). W przypadku wielu genów to właśnie poziom transkryptu mRNA decyduje o poziomie produktu białkowego. Ponadto geny wydają się bardziej różnić pomiędzy sobą na poziomie regulacji syntezy mRNA – tj. szybkości transkrypcji, niż na poziomie jego degradacji, choć dalsze badania procesu interferencji RNA mogą doprowadzić do zmiany tego poglądu.

Drugim czynnikiem wpływającym na rozwój badań nad współregulacją ekspresji genów na poziomie mRNA, stało się odczytanie sekwencji rosnącej liczby genomów, orga-

nizmów takich jak drożdże, człowiek, mysz, szympanś i inne. Odczytanie pełnych sekwencji genomów (a nie tylko sekwencji kodujących białka) pokazało, że większość genomu nie koduje białek i pozwoliło na globalną identyfikację prawdopodobnych sekwencji *cis*-regulatorowych. W przypadku złożonych organizmów takich jak kręgowce, obszary *cis*-regulatorowe zajmują większą część genomu niż (badane od kilku dziesięcioleci) obszary kodujące białka (BIRNEY i współaut. 2007). Przedrostek *cis*- (mający korzenie w klasycznych doświadczeniach genetycznych nad komplementacją) oznacza działanie fragmentu chromosomu (tu – motywu lub obszaru) na geny położone w jego sąsiedztwie na chromosomie, w odróżnieniu od działania *in trans* – wpływu na geny o odległych lokalizacjach na tym samym lub innych chromosomach. To wzajemne powiązanie (choć nie jeden-do-jednego) motywów *cis*-regulatorowych i czynników transkrypcyjnych umoż-

liwia obliczeniowe wnioskowanie o mechanizmach regulacji transkrypcji, na podstawie danych o ekspresji i sekwencji genomowej.

Trzecim składnikiem postępu w badaniach nad współregulacją jest doświadczalna identyfikacja miejsc wiążących dla setek czynników transkrypcyjnych oraz rozwój metod doświadczalnej identyfikacji miejsc wiązania białek do DNA w skali całego genomu. Uwzględnienie danych o aktywności czynników transkrypcyjnych i ich wiązaniu w skali całego genomu, pozwala na stworzenie prawdopodobnych modeli współregulacji transkrypcyjnej. Ta część badań jest najbardziej zaawansowana w przypadku drożdży i między innymi dlatego, mechanizmy współregulacji wielu genów są tam najlepiej poznane (SIMON i współaut. 2001, LEE i współaut. 2002). To, w jakim stopniu wyniki uzyskane w przypadku drożdży można uogólnić na bardziej złożone organizmy, takie jak organizm człowieka, stanowi obecnie jedno z kluczowych pytań w badaniach współregulacji genów.

Kolejnym czynnikiem decydującym o szybkim postępie badań nad współregulacją są wyniki badań wskazujące na powszechność regulacji transkrypcyjnej genów na poziomie modyfikacji postranslacyjnych białek chromatyny (głównie histonów). Modyfikacje takie określa się powszechnie w piśmiennictwie mianem modyfikacji epigenetycznych (choć są pewne spory dotyczące takiego użycia tego terminu, nie samego faktu modyfikacji). Uwzględnienie wyników badań stanu modyfikacji chromatyny w skali całego genomu (HON i współaut. 2008) w modelach współregulacji ekspresji genów stanowi kolejne aktualne zagadnienie badawcze.

Badając – w skali całego genomu – obszary regulatorowe lub obszary zmodyfikowanej epigenetycznie chromatyny, rozważano dotąd głównie „liniową” pozycję genu na chromosomie. Wiadomo jednak, że w jądrze komórki w interfazie cyklu komórkowego chromosomy zajmują określone (jak się wydaje, niezbyt ściśle) obszary chromatyny. Ponadto wiadomo, że różne rejony jądra komórkowego różnią się stanem aktywności

transkrypcyjnej chromatyny, co odzwierciedla klasyczny podział na aktywną euchromatynę i skondensowaną heterochromatynę. W obrębie euchromatyny zaobserwowano rejony o zwiększonej aktywności transkrypcyjnej tzw. fabryki transkrypcyjne. Wydaje się wielce prawdopodobne, że lokalizacja genu w trójwymiarowej strukturze jądra komórkowego może mieć wpływ na jego ekspresję. Od dawna wiadomo, że lokalizacja genu w heterochromatynie skutkuje brakiem ekspresji (np. nieaktywny chromosom X). Jak dotąd, brak metod doświadczalnych pozwalających na lokalizację wielu (docelowo wszystkich) genów w jądrze komórkowym w interfazie. Rozwinięto jednak metody pozwalające na identyfikację grup genów w trójwymiarowym sąsiedztwie określonego genu. Ocena wagi wpływu trójwymiarowej lokalizacji w jądrze na ekspresję wielu genów to istotne i aktualne pytanie badawcze (FRASER i BICKMORE 2007).

Epigenetycznym mechanizmem regulacji transkrypcji jest rodzicielskie piętno genomowe, zwane też naznaczeniem genetycznym (ang. genomic imprinting) polegające na ograniczeniu ekspresji niektórych ssaczych genów tylko do jednego z rodzicielskich alleli. Obszary naznaczone genetycznie znajdują się często w skupieniach w domenach chromosomów wielkości milionów par zasad i mają kilka cech wspólnych. Na przykład wszystkie skupienia w obszarze kontrolującym proces piętnowania genomowego (ang. imprinting control region, ICR), są zwykle długości 1–5 kb i są silnie metylowane. Zazwyczaj w tym obszarze kodowane jest przynajmniej jedno niekodujące RNA (ang. noncoding RNA, ncRNA) ulegające ekspresji zwykle z allelu matczynego i regulujące ekspresję wielu genów. Pojawianie się ncRNA może wyciszać ekspresję licznych genów w obu kierunkach lub blokować wiązanie się białka CTCF (ang. CCCTC binding factor binding) do obszarów wyciszających transkrypcję (STUART i współaut. 2003, BARTOLOMEI i Bartolomei 2008).

BIOINFORMATYCZNA ANALIZA WSPÓŁREGULACJI

Podstawowe dla badań współregulacji jest pytanie: „Które geny są regulowane, w określonej sytuacji biologicznej, w ten sam sposób?” Poznanie jednocześnie lub podobnie regulowanych grup genów pozwoliłoby na

taki wybór genów do szczegółowych badań doświadczalnych regulacji transkrypcji, aby ich wyniki można było rozszerzyć na grupę współregulowanych z nimi genów. Poznanie mechanizmów regulacji grup genów stanowi

warunek zrozumienia – na poziomie molekularnym – zachowania komórek jako systemu współdziałających elementów.

Badania współregulacji transkrypcji posługują się metodami porównawczymi: porównania odpowiadających sobie genów (zwłaszcza tzw. genów ortologicznych, odpowiadających sobie jeden do jednego) pomiędzy genomami różnych gatunków pozwalają na identyfikację przypuszczalnych obszarów regulatorowych. Łączna analiza przypuszczalnych sekwencji regulatorowych i wzorców ekspresji wielu genów pozwala na identyfikację cech sekwencji DNA odgrywających rolę w tworzeniu określonych wzorców ekspresji. To zadanie, określane mianem analizy regulacji transkrypcyjnej *in silico*, stanowi jeden z najciekawszych wyzwań współczesnej biologii.

Jak wynika z zarysowanego powyżej stanu badań doświadczalnych, grupa genów może podlegać współregulacji transkrypcyjnej na kilku poziomach: wiązania czynników transkrypcyjnych i inicjacji transkrypcji, modyfikacji chromatyny, lokalizacji jądrowej, zmiany

stanu upakowania chromatyny, przy czym relatywna waga wymienionych powyżej poziomów regulacji rośnie od lewej do prawej wraz z wydłużaniem skali czasowej lub wzrostem siły bodźca.

Dane z doświadczeń wielkoskalowych (ang. high-throughput) używane dotąd do analizy współregulacji ekspresji, to dane z profilowania ekspresji oraz dane o sekwencji genomowej. Te ostatnie są zwykle analizowane z wykorzystaniem bibliotek znanych motywów (sekwencji DNA o długości kilku, kilkunastu lub kilkudziesięciu nukleotydów) miejsc wiążących czynniki transkrypcyjne. Oznacza to, że w większości dotychczasowych analizach regulacji transkrypcyjnej *in silico* nie uwzględniano stanu modyfikacji epigenetycznej chromatyny. Stan upakowania chromatyny był pośrednio częściowo uwzględniany, jako że geny w heterochromatinie nie ulegają ekspresji. Jednak maszyneria modyfikująca chromatynę wiąże się do DNA za pośrednictwem czynników transkrypcyjnych, co pozostawia ślad w sekwencji – dostępny dla metod bioinformatycznych.

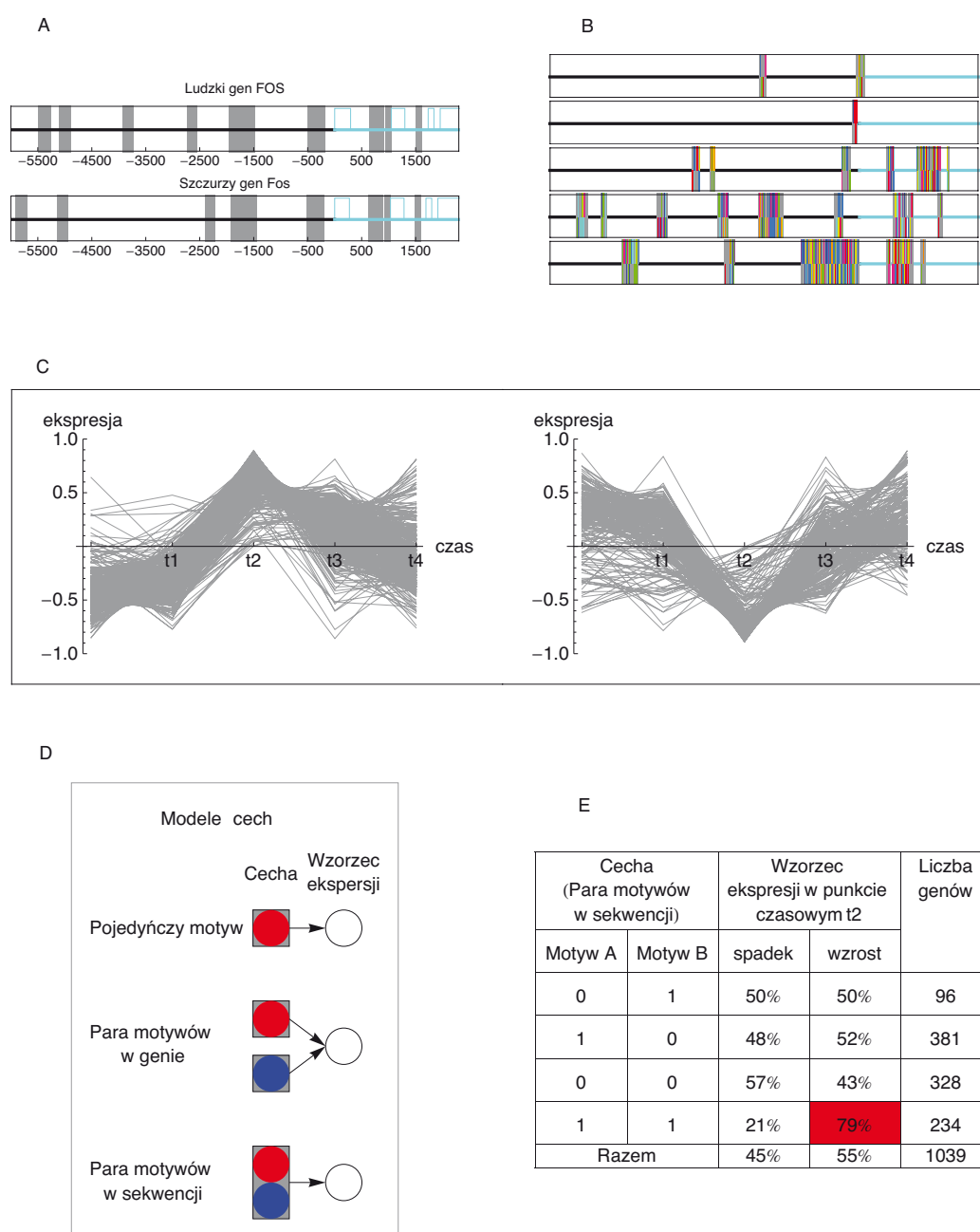
METODY POSZUKIWANIA OBSZARÓW CIS-REGULATOROWYCH

Analizę bioinformatyczną regulacji transkrypcyjnej można rozpatrywać jako kilka wzajemnie związanych zagadnień. Ważnym etapem analizy *in silico* regulacji transkrypcyjnej jest identyfikacja przypuszczalnych obszarów (sekwencji) *cis*-regulatorowych genów. Ten etap jest konieczny i kluczowy dla sukcesu dalszych etapów obliczeniowej analizy *cis*-regulacji, ze względu na ograniczenia czułości i specyficzności metod bioinformatycznych identyfikacji motywów *cis*-regulatorowych. Zadanie identyfikacji przypuszczalnym obszarów regulatorowych jest stosunkowo proste u drożdży, gdzie zwykle są one stosunkowo krótkie (rzędu 500 par zasad) i położone w sąsiedztwie obszarów kodujących, bezpośrednio poniżej miejsca startu transkrypcji – w przypadku genów drożdży mamy do czynienia tylko z promotorami.

W przypadku organizmów tkankowych (zwłaszcza u kręgowców), sytuacja jest bardziej skomplikowana. Na podstawie wyników badań doświadczalnych uzyskanych metodami biochemicznymi wiadomo, że obszary regulatorowe kręgowców są znacznie dłuższe, i (podobnie jak obszary kodujące) złożone z wielu rozdzielonych fragmentów, z których jedynie promotor ma w przybli-

żeniu stałą lokalizację poniżej (5') obszaru kodującego, a pozostałe, określane jako sekwencje wzmacniające, lub obszary sterujące (ang. locus control region), mają różne lokalizacje, niekiedy w znacznej odległości od regulowanego genu. Do niedawna badania doświadczalne przeważnie ograniczały się do badań promotorów i w ten sposób zidentyfikowano doświadczalnie setki promotorów, których sekwencje umożliwiły opracowania pierwszej, specyficznej metody bioinformatycznej identyfikacji promotorów u kręgowców. Metoda oparta jest o analizę częstości występowania wszystkich możliwych par krótkich motywów, przedzielonych krótkim łącznikiem.

Jednak przełomem, jak już wspomniano, stało się odczytanie pełnych sekwencji kilku genomów kręgowców, co dzięki zastosowaniu rozwiniętych wcześniej metod bioinformatycznych wzajemnego dopasowania (uliniowienia, ang. alignment) sekwencji, pozwoliło na identyfikację tzw. ewolucyjnie zachowanych obszarów niekodujących, we flankach genomowych tysięcy odpowiadających sobie genów (Ryc. 2) (BRAY i współaut. 2003, BRUDNO i współaut. 2003). Wyniki weryfikacji doświadczalnych wskazują, że znacz-



Ryc. 2. Identyfikacja cech obszarów regulatorowych związanych z określonym wzorcem ekspresji.

A. Przypuszczalne rejony cis-regulatorowe (zachowane ewolucyjnie sekwencje niekodujące), zidentyfikowane przy użyciu programów AVID-VISTA (BRAY i współaut. 2003), dla ortologicznych genów człowieka i szczura, zaznaczone jako szare prostokąty. Na seledynowo zaznaczono region kodujący i eksony. B. Miejsca wiążące znane czynniki transkrypcyjne (zaznaczone różnymi kolorami) w przypuszczalnych rejonach cis-regulatorowych pięciu różnych genów człowieka, zidentyfikowane z zastosowaniem komercyjnej biblioteki motywów. C. Arbitralny podział genów na dwie grupy względem prostego wzorca ekspresji: wzrost lub spadek ekspresji w punkcie czasowym t2 w stosunku do t0. Przykładowe wyniki, w doświadczalnym modelu udaru, uzyskane we współpracy z dr Małgorzatą Zawadzką, IBD PAN im. M. Nenckiego. D. Przykłady prostych cech cis-regulatorowych, reprezentowanych jako sieci zależności. Para motywów w genie nie musi wystąpić w obrębie tej samej sekwencji cis-regulatorowej tego genu. E. Tabela ilustrująca statystyczny związek pomiędzy współwystąpieniem pary motywów w genie a określonym wzorcem ekspresji. Przykładowe wyniki uzyskane we współpracy z dr Norbertem Dojerem (Instytut Informatyki Uniwersytetu Warszawskiego).

ny odsetek tak zidentyfikowanych sekwencji to prawdziwe obszary wzmacniające ekspresję (PENNACCHIO i współaut. 2006, VISEL i współaut. 2008).

Rozwijane są alternatywne metody bioinformatyczne identyfikacji obszarów wzmacniających ekspresję (lub ich podjednostek, zwanych modułami *cis*-regulatorowymi). Metody te oparte są o dopasowanie sekwencji nukleotydowych występujących w danym module znanych motywów wiążących czynniki transkrypcyjne z uwzględnieniem kolejności wystąpień tych motywów (HALLIKAS i współaut. 2006). W tak zidentyfikowanych obszarach wzmacniających (ang. enhancers), odnajduje się wystąpienia tego samego motywu (co najmniej dwukrotne i zachowane pomiędzy kilkoma gatunkami) w określonych odległościach, jako cechy pozwalające na predykcję określonego wzorca ekspresji genów ze zweryfikowaną doświadczalnie specyficznością rzędu 30%.

Niedawne sukcesy predykcji obszarów regulatorowych metodami bioinformatycznymi mogą zostać przyćmione przez niezwykle szybki postęp wielkoskalowych metod doświadczalnych, w szczególności immunoprecypitacji chromatyny w skali całego genomu (HEINTZMAN i REN 2007). Przeprowadzone tą techniką badania wiązania do chromosomów jednego z białek modyfikujących chromatynę – acetylotransferazy histonów Ep300, wskazują, że identyfikuje ono obszary wzmacniające ekspresję aktywne w danej tkance. Można się spodziewać, że w oparciu o nowe dane

doświadczalne powstaną wkrótce modele pozwalające przewidzieć położenie obszarów wzmacniających ekspresję lub modułów regulatorowych bez odwoływania się do zachowania ich sekwencji pomiędzy gatunkami. Co ważniejsze, takie modele – po ich weryfikacji doświadczalnej – pogłębią nasze zrozumienie o wewnętrznej organizacji, a co za tym idzie funkcjonowaniu obszarów wzmacniających ekspresję.

Problemem, który do niedawna nie miał zadowalającego rozwiązania, a który może znaleźć rozwiązanie doświadczalne, jest przypisanie obszarów wzmacniających ekspresję do genów. W publikacjach bioinformatycznych zwykle arbitralnie przypisywano takie obszary do jednego lub kilku najbliższych położonych genów, podczas gdy wyniki doświadczalne uzyskane metodami biochemicznymi wskazują na możliwość ich oddziaływania na odległość rzędu miliona par zasad. Wyniki najnowszych badań doświadczalnych wskazują, że obszary wzmacniające ekspresję oddziałują na geny na tym samym chromosomie nie oddzielone od nich specjalnymi sekwencjami *cis*-regulatorowymi, tzw. sekwencjami wyciszającymi (ang. silencers lub insulators), których funkcją jest właśnie izolacja od wpływu obszarów wzmacniających (CHAN i SONG 2008). Zidentyfikowano CTCF (ang. CCCTC binding factor) jako białko specyficznie wiążące się z sekwencjami wyciszającymi oraz doświadczalnie ustalono miejsca jego wiązania do genomu.

METODY POSZUKIWANIA WSPÓLREGULOWANYCH GENÓW

Identyfikacji genów, o których przypuszcza się, że są współregulowane, można dokonać na wiele sposobów, podejmując przy tym dwa zasadnicze wybory. Pierwszy zasadniczy wybór polega na tym, czy analizujemy dane w układzie współrzędnych pomiaru (np. w doświadczeniu kinetycznym, gdy badamy ekspresję w kolejnych punktach czasowych), czy po ich przekształceniu do innego układu współrzędnych, który może (ale nie musi) ułatwiać dalszą analizę. Na przykład, niemal powszechnie używane jest przekształcenie danych do postaci logarytmicznej, co pozwala traktować symetrycznie wzrosty i spadki ekspresji. Często stosowane są zabiegi normalizacji lub standaryzacji, pozwalające wydobyć podobieństwa wzorców o podobnym kształcie, lecz różnej sile zmian ekspre-

sji, jednak kosztem utraty informacji o tych ostatnich. Na przykład w doświadczeniu porównującym dwa warunki eksperymentalne, niewielki wzrost i duży wzrost ekspresji, o ile oba wzrosty są statystycznie znamienne, będą traktowane tak samo. Stosunkowo często stosowana bywa ortogonalizacja (w statystyce nazywana analizą składowych głównych), czyli takie przekształcenie danych do takiego układu współrzędnych, że ich wartości są wzajemnie nieskorelowane (HOLTER i współaut. 2000). Inne stosowane metody to rozkład na składowe niezależne (LIEBERMEISTER 2002), rozkład na składowe nieujemne (THOMPSON i współaut. 2008), itd.

Drugi zasadniczy wybór to albo identyfikacja genów przypuszczalnie współregulowanych w oparciu o całość dostępnych danych

o ekspresji, albo poszukiwanie genów współregulowanych w pewnej części przestrzeni pomiaru (np. w określonym punkcie czasowym, albo wzdłuż wybranych osi współrzędnych w przekształconych danych).

Wybór analizy wielowymiarowych danych (np. wiele punktów czasowych lub warunków doświadczenia, lub równocześnie analizowanych składowych głównych) pociąga zwykle za sobą konieczność pogrupowania genów w grupy o podobnej ekspresji za pomocą analizy skupień, niekiedy utożsamianej z klasteryzacją. Istnieje wiele algorytmów analizy skupień, spośród których najpowszechniej używane w transkryptomice to metoda klasteryzacji hierarchicznej i metoda K-średnich (DO i CHOI 2008).

Analiza współregulacji w określonym punkcie czasowym (BUSSEMAKER i współaut. 2001), albo (bardziej ogólnie) wzdłuż wybranych osi współrzędnych w przekształconych danych (rozkład na składowe), również pozwalają na podzielenie genów na grupy o różnej ekspresji, także w sytuacji gdy geny

nie tworzą wyraźnie odrębnych skupień o różnych wzorcach ekspresji. W przypadku rozkładu na składowe, np. składowe główne, celowe jest wykazanie, że składowe wybrane do dalszej analizy odzwierciedlają biologiczne mechanizmy regulacji. Zaproponowaliśmy w tym celu metodę polegającą na identyfikacji składowych zachowanych pomiędzy wynikami profilowania ekspresji w niezależnych doświadczeniach w powiązanych układach doświadczalnych (np. tkanka nerwowa *in vivo* i neurony w hodowli) (DĄBROWSKI i współaut. 2006a).

W przypadku składowych zachowanych można się spodziewać, że różne składowe odzwierciedlają aktywność różnych regulatorów (ALTER i współaut. 2000), a więc, że regulacja składowych jest prostsza niż – będącego ich sumą – pełnego profilu ekspresji we wszystkich warunkach pomiaru. Prostsza regulacja powinna się przekładać na prostsze cechy sekwencji regulatorowych, co zwiększa szanse ich identyfikacji w oparciu o dane o ekspresji.

OD WZORÓW EKSPRESJI GENÓW DO MOTYWÓW REGULATOROWYCH

Najprostszymi i najczęściej rozważanymi w tym kontekście cechami są krótkie (kilkunastonukleotydowe) sekwencje – tzw. motywy sekwencji będące miejscem wiązania czynników transkrypcyjnych lub innych regulatorów transkrypcji. W badaniach współregulacji wykorzystywane są dwie zasadnicze metody identyfikacji motywów w sekwencjach regulatorowych: identyfikacja *de novo* – na podstawie zbioru potencjalnych sekwencji *cis*-regulatorowych przypuszczalnie współregulowanych genów (HUGHES i współaut. 2000) oraz identyfikacja wcześniej znanych (zidentyfikowanych doświadczalnie) motywów wiążących czynniki transkrypcyjne. W identyfikacji miejsc wiązania czynników transkrypcyjnych stosuje się programy wyszukujące tzw. miejsca konsensusowe z wykorzystaniem wyspecjalizowanych baz danych, tzw. bibliotek motywów (MATYS i współaut. 2006, BRYNE i współaut. 2008). „Konsensus” to określona sekwencja nukleotydowa wiążąca *in vitro* dany czynnik transkrypcyjny, najczęściej występująca w promotorach znanych genów. Często jest to sekwencja zweryfikowana doświadczalnie, o dobrze scharakteryzowanych oddziaływaniach molekularnych białko-DNA. Obecnie najczęściej stosuje się wyznaczanie sekwencji

z uwzględnieniem pozycyjnych macierzy częstości (ang. position weight matrix, PWM), co w luźnym tłumaczeniu oznacza „siłę pozycji w macierzy” i pozwala przypisać większą wagę mniej zmiennym pozycjom. Istnieją też ciekawe podejścia, polegające na łączeniu przewidywań *de novo*, bibliotek motywów oraz uwzględniające zachowanie motywów pomiędzy odpowiadającymi sobie pozycjami w genomie u różnych gatunków (LOOTS i OVCHARENKO 2004), a nawet zależności ewolucyjne tych ostatnich.

Opisane powyżej zasadnicze wybory dotyczące sposobu identyfikacji genów potencjalnie współregulowanych, mają wpływ na ogólne modele i sposób poszukiwania bardziej złożonych niż pojedyncze motywy cech odpowiedzialnych za współregulację. Przykładowo, podział zbioru danych na klastry, powoduje, że możemy szukać cech sekwencyjnych (jakościowych lub ilościowych) związanych z przynależnością do określonego klastra. W piśmiennictwie dominuje model, w którym cechą decydującą o przynależności genu do klastra o określonym wzorcu ekspresji jest obecność w jego promotorze układu motywów wiążących kilka różnych czynników transkrypcyjnych, być może z dodatkowymi ograniczeniami dotyczącymi ich krotności,

położenia, itp. W przełomowej pracy, opisującej pierwszą udaną predykcję ekspresji wielu genów na podstawie sekwencji genomowej nicienia *C. elegans*, BEER i TAVAZOIE (2004) zastosowali formalizm sieci bayesowskich, wiążących prawdopodobieństwo przynależności genu do określonego skupienia z cechami (motywami, odległościami, itd.) określonego układu motywów w promotorach genów.

W sytuacji doświadczalnej, gdy geny nie tworzą wyraźnie odrębnych skupień o różnym wzorze ekspresji, celowe może być zastosowanie modelu, w którym zmiana ekspresji modelowana jest jako ciągła funkcja wartości pewnego zbioru cech sekwencji regulatorowych, w szczególności rodzaju i liczby motywów wiążących czynniki transkrypcyjne. Taki liniowy model odpowiedzi genów znajduje uzasadnienie w klasycznych wynikach

badania nad wpływem krotności miejsc wiążących na poziom ekspresji genów (LAYBOURN i KADONAGA 1992), w fizykochemicznym opisie przyłączania maszynerii transkrypcyjnej do promotora. Zarysowany powyżej opis modelu liniowego dotyczy stanów komórki bliskich równowagi. W stanach odległych od równowagi (np. w doświadczeniach perturbacyjnych), model liniowy formułowany jest w języku równań różniczkowych. Dokładniejszą charakterystykę modelu liniowego można znaleźć w przystępnej pracy przeglądowej (BUSSEMAKER i współaut. 2007).

Wyniki niezwykle szczegółowych badań promotora genu *endo16* jeżowca (DAVIDSON 2006) sugerują, że oba modele (moduły *cis*-regulatorowe i model log-liniowy) są częściowo prawdziwe, a zarazem, że nie wyczerpują one rzeczywistej złożoności organizacji i funkcji obszarów regulatorowych.

METODY WERYFIKACJI PRZEWIDYWAŃ BIOINFORMATYCZNYCH

Kluczowe jest pytanie o możliwość weryfikacji wyników przewidywań bioinformatycznych o roli określonych cech sekwencyjnych w determinowaniu określonych wzorców ekspresji. Warto podkreślić, że możliwa i celowa jest weryfikacja *in silico* przewidywań uzyskanych metodami bioinformatycznymi, w oparciu o dane dotyczące innych genów niż te użyte do generowania przewidywań. Istnieje wiele metod, spośród których najprostsza, choć wymagająca największej ilości danych, polega na podziale wyjściowego zbioru danych (danych o ekspresji i o sekwencji genomowej) losowo na dwie części. Na jednym zbiorze danych, określanym jako zbiór uczący, identyfikujemy cechy sekwencji związane z określonym wzorcem ekspresji. Następnie cech zidentyfikowanych w zbiorze uczącym używamy, aby przewidzieć ekspresję innych genów zawierających te same cechy w zbiorze testowym (BEER i TAVAZOIE 2004). Możliwa jest również weryfikacja zidentyfikowanych cech *cis*-regulatorowych tych samych genów, w innych warunkach doświadczenia profilowa-

nia ekspresji. Stosując to podejście, pomyślnie zweryfikowaliśmy motywy wiążące czynniki transkrypcyjne E2F i NR2F, jako cechy związane ze zmianami ekspresji genów na wczesnym etapie różnicowania neuronów u szczura (DĄBROWSKI i współaut. 2006b).

Opisane powyżej prace stanowią raczej wyjątek niż regułę, gdyż liczba prac, w których udało się określić i zweryfikować przewidywania o roli określonych cech sekwencyjnych w regulacji ekspresji wielu genów u kręgowców, pozostaje stosunkowo niewielka. Należy się jednak spodziewać, że uwzględnienie w modelach nowych danych doświadczalnych o pozycjach aktywnych obszarów wzmacniających lub wyciszających ekspresję oraz stale rosnące ilości i jakości danych o wzorcach ekspresji, przyczynią się wkrótce do znacznego wzrostu jakości przewidywań ekspresji genów u kręgowców na podstawie ich sekwencji.

Praca wykonana w ramach realizacji projektu PBZ-MNiI-2/1/2005.

PREDICTING AND MODELING GENE EXPRESSION FROM GENOMIC SEQUENCES – HOPES AND CHALLENGES

Summary

Control of eukaryotic transcription involves combinatorial interactions between transcription factors and regulatory sequences in the genome. In addition,

chromatin structure and modification states play key roles in the complex control of transcription. The maturation of methods for detecting *in vivo* protein-

DNA interactions on a genome-wide scale, detailed maps of chromatin features and transcription factor localization over entire genomes of eukaryotic cells increases our understanding of the properties and nature of transcriptional regulatory mechanisms. The knowledge gained from large-scale investigations has challenged some conventional concepts of promoter structure and function, and provided valuable insights into the complex gene regulatory mechanisms. Here, we review recent experimental and computational advances that have enabled identification and analysis of regulatory regions on an

unprecedented scale, allowing for systematic determination of the transcriptional regulatory networks in eukaryotic cells. Analysis of high-throughput gene expression data together with genomic sequence(s) holds promise of revealing features of cis-regulatory sequences responsible for co-regulation of many genes. This general approach has proved successful in simpler organisms, and recent studies in Metazoa reported features over-represented in groups of genes with particular expression patterns revealing gene regulatory mechanisms.

LITERATURA

- AERTS S., VAN LOO P., THIJS G., MOREAU Y., DE MOOR B., 2003. *Computational detection of cis-regulatory modules*. Bioinformatics 19 (Suppl. 2), 5–14.
- ALTER O., BROWN P. O., BOTSTEIN D., 2000. *Singular value decomposition for genome-wide expression data processing and modeling*. Proc. Natl. Acad. Sci. USA 97, 10101–10106.
- BARRETT T., TROUP D. B., WILHITE S. E., LEDOUX P., RUDNEV D., EVANGELISTA C., KIM I. F., SOBOLEVA A., TOMASHEVSKY M., MARSHALL K. A., PHILLIPPY K. H., SHERMAN P. M., MUERTTER R. N., EDGAR R., 2009. *NCBI GEO: archive for high-throughput functional genomic data*. Nucleic Acids Res. 37, D885–D890.
- BEER M. A., TAVAZOIE S., 2004. *Predicting gene expression from sequence*. Cell 117, 185–198.
- BIRNEY E., STAMATOYANNOPOULOS J. A., DUTTA A., GUGLIO R., GINGERAS T. R., MARGULIES E. H., WENG Z., SNYDER M., DERMITZAKIS E. T., THURMAN R. E., KUEHN M. S., TAYLOR C. M., NEPH S., KOCH C. M., ASTHANA S., MALHOTRA A., ADZHUBEI I., GREENBAUM J. A., ANDREWS R. M., FLICEK P. i współaut., 2007. *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project*. Nature 447, 799–816.
- BRAY N., DUBCHAK I., PACHTER L., 2003. *AVID: A global alignment program*. Genome Res. 13, 97–102.
- BRUDNO M., DO C. B., COOPER G. M., KIM M. F., DAVYDOV E., GREEN E. D., SIDOW A., BATZOGLOU S., 2003. *LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA*. Genome Res. 13, 721–731.
- BRYNE J. C., VALEN E., TANG M. H., MARSTRAND T., WINTHER O., DA PIEDADE I., KROGH A., LENHARD B., SANDELIN A., 2008. *JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update*. Nucleic Acids Res. 36, D102–D106.
- BUSSEMAKER H. J., LI H., SIGGIA E. D., 2001. *Regulatory element detection using correlation with expression*. Nature Genet. 27, 167–171.
- BUSSEMAKER H. J., FOAT B. C., WARD L. D., 2007. *Predictive modeling of genome-wide mRNA expression: from modules to molecules*. Ann. Rev. Biochem. Biomol. Struct. 36, 329–347.
- CHAN C. S., SONG J. S., 2008. *CCCTC-binding factor confines the distal action of estrogen receptor*. Cancer Res. 68, 9041–9049.
- DĄBROWSKI M., ADACH A., AERTS S., MOREAU Y., KAMIŃSKA B., 2006a. *Identification of conserved modes of expression profiles during hippocampal development and neuronal differentiation in vitro*. J. Neurochem. 97, Suppl. 1., 87–91.
- DĄBROWSKI M., AERTS S., KAMIŃSKA B., 2006b. *Prediction of a key role of motifs binding E2F and NR2F in down-regulation of numerous genes during the development of the mouse hippocampus*. BMC bioinformatics 7, 367.
- DAVIDSON E. H., 2006. *The Regulatory Genome*. San Diego, Academic Press.
- DO J. H., CHOI D. K., 2008. *Clustering approaches to identifying gene expression patterns from DNA microarray data*. Mol. Cells 25, 279–288.
- FRASER P., BICKMORE W., 2007. *Nuclear organization of the genome and the potential for gene regulation*. Nature 447, 413–417.
- GROSS P., OELGESCHLAGER T., 2006. *Core promoter-selective RNA polymerase II transcription*. Biochem. Soc. Symp. 73, 225–236.
- GUPTA M., LIU J. S., 2005. *De novo cis-regulatory module elicitation for eukaryotic genomes*. Proc. Natl. Acad. Sci. USA 102, 7079–7084.
- HALLIKAS O., PALIN K., SINJUSHINA N., RAUTIAINEN R., PARTANEN J., UKKONEN E., TAIPALE J., 2006. *Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity*. Cell 124, 47–59.
- HARBISON C. T., GORDON D. B., LEE T. I., RINALDI N. J., MACISAAC K. D., DANFORD T. W., HANNETT N. M., TAGNE J. B., REYNOLDS D. B., YOO J., JENNINGS E. G., ZEITLINGER J., POKHOLOK D. K., KELLIS M., ROFFE P. A., TAKUSAGAWA K. T., LANDER E. S., GIFFORD D. K., FRAENKEL E., YOUNG R. A., 2004. *Transcriptional regulatory code of a eukaryotic genome*. Nature 431, 99–104.
- HEINTZMAN N. D., REN B., 2007. *The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome*. Cell Mol. Life Sci. 64, 386–400.
- HOLTER N. S., MITRA M., MARITAN A., CIEPLAK M., BANAVAR J. R., FEDOROFF N. V., 2000. *Fundamental patterns underlying gene expression profiles: simplicity from complexity*. Proc. Natl. Acad. Sci. USA 97, 8409–8414.
- HON G., REN B., WANG W., 2008. *ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome*. PLoS Computat. Biol. 4, e1000201.
- HUGHES J. D., ESTEP P. W., TAVAZOIE S., CHURCH G. M., 2000. *Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae*. J. Mol. Biol. 296, 1205–1214.
- KUMMERFELD S. K., TEICHMANN S. A., 2006. *DBD: a transcription factor prediction database*. Nucleic Acids Res. 34, D74–D81.
- LAYBOURN P. J., KADONAGA J. T., 1992. *Threshold phenomena and long-distance activation of transcription by RNA polymerase II*. Science 257, 1682–1685.
- LEE T. I., RINALDI N. J., ROBERT F., ODOM D. T., BAR-JOSEPH Z., GERBER G. K., HANNETT N. M., HARBISON C. T., THOMPSON C. M., SIMON I., ZEITLINGER J., JENNINGS E. G., MURRAY H. L., GORDON D. B.,

- REN B., WYRICK J. J., TAGNE J. B., VOLKERT T. L., FRAENKEL E., GIFFORD D. K. i współaut., 2002. *Transcriptional regulatory networks in Saccharomyces cerevisiae*. Science 298, 799–804.
- LEVINE M., TJIAN R., 2003. *Transcription regulation and animal diversity*. Nature 424, 147–151.
- LIEBERMEISTER W., 2002. *Linear modes of gene expression determined by independent component analysis*. Bioinformatics 18, 51–60.
- LOOTS G. G., OVCHARENKO I., 2004. *rVISTA 2.0: evolutionary analysis of transcription factor binding sites*. Nucleic Acids Res. 32, W217–W221.
- MATYS V., KEL-MARGOULIS O. V., FRICKE E., LIEBICH I., LAND S., BARRE-DIRRIE A., REUTER I., CHEKMENEV D., KRULL M., HORNISCHER K., VOSS N., STEGMAIER P., LEWICKI-POTAPOV B., SAXEL H., KEL A. E., WINGENDER E., 2006. *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes*. Nucleic Acids Res. 34, D108–D110.
- ORPHANIDES G., REINBERG D., 2002. *A unified theory of gene expression*. Cell 108, 439–451.
- PARKINSON H., KAPUSHESKY M., KOLESNIKOV N., RUSTICI G., SHOJATALAB M., ABEGUNAWARDENA N., BERUBE H., DYLAG M., EMAM I., FARNE A., HOLLOWAY E., LUKK M., MALONE J., MANI R., PILICHEVA E., RAYNER T. F., REZWAN F., SHARMA A., WILLIAMS E., BRADLEY X. Z. i współaut., 2009. *ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression*. Nucleic Acids Res. 37, D868–D872.
- PENNACCHIO L. A., AHITUV N., MOSES A. M., PRABHAKAR S., NOBREGA M. A., SHOUKRY M., MINOVITSKY S., DUBCHAK I., HOLT A., LEWIS K. D., PLAJZER-FRICK I., AKIYAMA J., DE VAL S., AFZAL V., BLACK B. L., COURONNE O., EISEN M. B., VISEL A., RUBIN E. M., 2006. *In vivo enhancer analysis of human conserved non-coding sequences*. Nature 444, 499–502.
- SEGAL E., RAVEH-SADKA T., SCHROEDER M., UNNERSTALL U., GAUL U., 2008. *Predicting expression patterns from regulatory sequence in Drosophila segmentation*. Nature 451, 535–540.
- SHARAN R., OVCHARENKO I., BEN-HUR A., KARP R. M., 2003. *CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments*. Bioinformatics 19, Suppl. 1, i283–i291.
- SIMON I., BARNETT J., HANNETT N., HARBISON C. T., RINALDI N. J., VOLKERT T. L., WYRICK J. J., ZEITLINGER J., GIFFORD D. K., JAAKKOLA T. S., YOUNG R. A., 2001. *Serial regulation of transcriptional regulators in the yeast cell cycle*. Cell 106, 697–708.
- SMALE S. T., KADONAGA J. T., 2003. *The RNA polymerase II core promoter*. An. Rev. Biochem. 72, 449–479.
- STUART J. M., SEGAL E., KOLLER D., KIM S. K., 2003. *A gene-coexpression network for global discovery of conserved genetic modules*. Science 302, 249–255.
- THOMPSON C. L., PATHAK S. D., JEROMIN A., NG L. L., MACPHERSON C. R., MORTRUD M. T., CUSICK A., RILEY Z. L., SUNKIN S. M., BERNARD A., PUCHALSKI R. B., GAGE F. H., JONES A. R., BAJIC V. B., HAWRYLYCZ M. J., LEIN E. S., 2008. *Genomic anatomy of the hippocampus*. Neuron 60, 1010–1021.
- VISEL A., PRABHAKAR S., AKIYAMA J. A., SHOUKRY M., LEWIS K. D., HOLT A., PLAJZER-FRICK I., AFZAL V., RUBIN E. M., PENNACCHIO L. A., 2008. *Ultraconservation identifies a small subset of extremely constrained developmental enhancers*. Nat. Genet. 40, 158–160.
- BARTOLOMEI L. B., BARTOLOMEI M. S., 2008. *Regulation of imprinting in clusters: noncoding RNAs versus insulators*. Adv. Genet. 61, 207–223.