Metody macierzowe w analizie danych transkryptomicznych i metabolomicznych

Autoreferat

Krzysztof Gogolewski

Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski

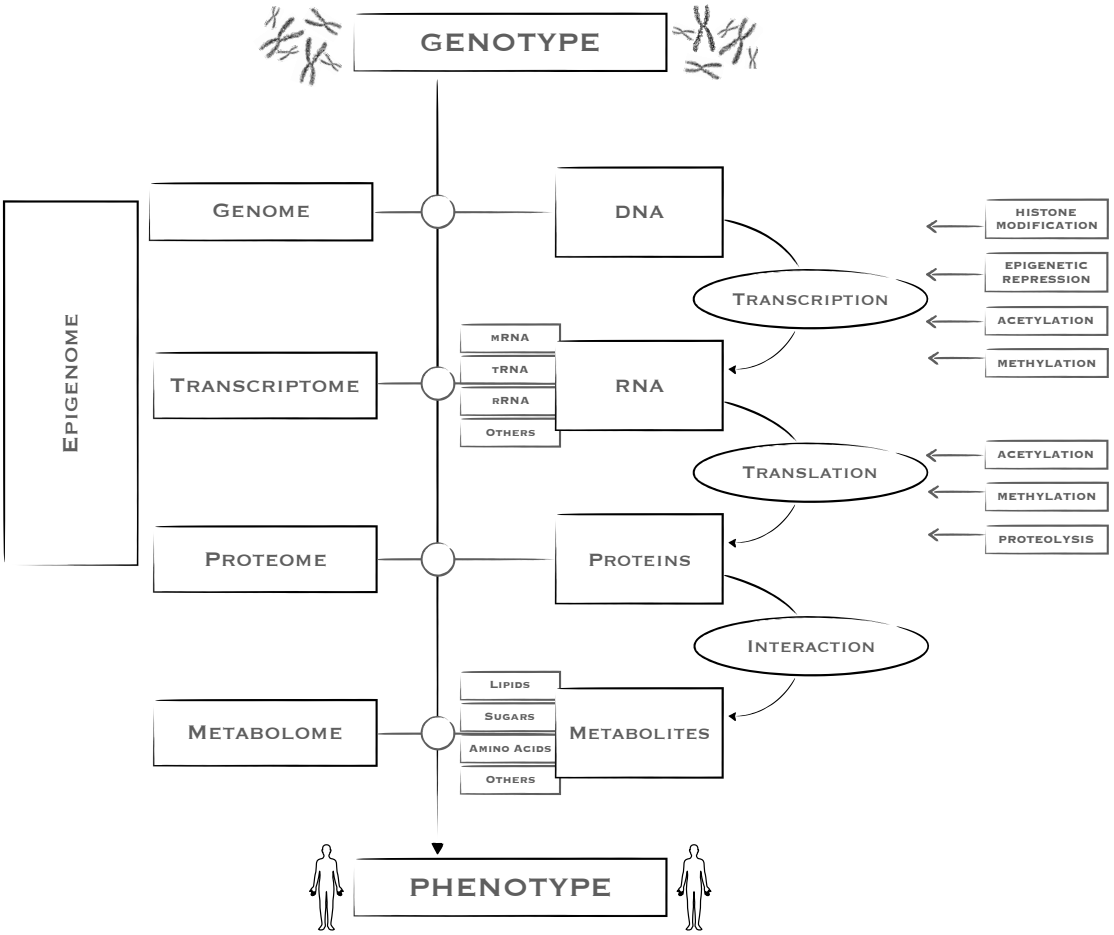
1 Wstęp

Intensywny rozwój technologiczny i naukowy w dziedzinie biotechnologii, w sposób naturalny pociąga za sobą konieczność opracowywania nowych bioinformatycznych metod przeznaczonych do analizy szybko przyrastających zasobów danych biomedycznych i molekularnych. Dane te gromadzone są zazwyczaj dzięki zastosowaniu, tzw. technologii wysokoprzepustowych [Reuter et al., 2015], które pozwalają na wydajne przeprowadzanie eksperymentów na dużą skalę, np. skalę całego genomu.

W dziedzinie bioinformatyki i biologii systemów, technologie te mają na celu prowadzenie precyzyjnych badań w zakresie biologii komórki zmierzających do określenia jej obrazu molekularnego za pomocą: struktury genomu (genomika, ang. *genomics*), pomiarów transkryptomicznych, tj. poziomów matrycowego RNA w komórce (transkryptomika, ang. *transcriptomics*), struktury i funkcji białek (proteomika, ang. *proteomics*), zachodzących procesów metabolomicznych (metabolomika, ang. *metabolomics*) i innych (patrz Rysunek 1). W nawiasach podane zostały nazwy dziedzin, które zajmują się analizą odpowiedniego rodzaju danych. Wszystkie te dziedziny stanowią rodzinę nauk nazywaną *-omics studies*. Mają one na celu przetwarzanie i charakteryzowanie dostępnych danych wielkoskalowych oraz wnioskowanie na ich podstawie o strukturze, funkcji i dynamice pojedynczych komórek, organizmów czy też całych populacji [Hasin et al., 2017].

Spośród powyżej wymienionych obszarów naukowych, w rozprawie szczególna uwaga zostaje poświęcona transkryptomice [Lowe et al., 2017] i metabolomice [Riekeberg and Powers, 2017] oraz odpowiadających im typom danych. Obie te dziedziny opierają się na wnioskowaniu z wysokoprzepustowych danych molekularnych, w badaniach prowadzonych w zakresie biologii ogólnej [Shin et al., 2018], medycyny spersonalizowanej [Li et al., 2016], diagnostyki nowotworów [Ren et al., 2016] czy ich klasyfikacji [Borgan et al., 2010], jak również poznawania etiologii chorób [Stempler et al., 2014, Borrageiro et al., 2018].

W rozprawie zostało przedstawionych i opisanych kilka nowych metod obliczeniowych oraz efektywnych algorytmów pozwalających na analizę danych transkryptomicznych i metabolomicznych. Dodatkowo, w każdym rozdziale zilustrowane zostało użycie poszczególnych metod do rozwiązania konkretnych problemów biomedycznych, w których niezbędna jest analiza danych pochodzących z wysokoprzepustowych technologii. Wobec powyższego, zasadniczym celem rozprawy jest rozwój metod i algorytmów, które przyczynią się do lepszego zrozumienia procesów molekularnych z zakresu biologii komórki, poprzez analizę dostępnych danych biomedycznych. Jednocześnie, kolejne rozdziały niniejszej pracy mają na celu wykazanie, że wartość naukowa przedstawionych w nich wyników stanowi istotny wkład w rozwój interdyscyplinarnego świata bioinformatyki.



Rysunek 1: **Przepływ informacji genetycznej w komórce.** Rysunek przedstawia kolejne etapy mapowania ludzkiego genotypu na jego fenotyp. Na każdym etapie przetwarzana jest informacja molekularna, prowadząca od kodu genetycznego zawartego w genotypie danej osoby, poprzez liczne procesy genetyczne i epigenetyczne, do obserwowalnych cech jednostki.

1.1 Mapowanie genotyp-fenotyp i motywacja

Wyzwaniem, dla którego dobrze jest poznać i zrozumieć wartość potencjału obliczeniowego i algorytmicznego informatyki w dziedzinie współczesnej biologii molekularnej jest zagadnienie mapowania genotypu na fenotyp.

Po raz pierwszy, w [Alberch, 1991] autorzy sugerują istnienie takiego mapowania, gdzie jako *genotyp* należy rozumieć pełną dziedziczną informację opisującą organizm zaś *fenotyp* stanowią obserwowalne właściwości danego organizmu, takie jak morfologia czy fizjologia. Od czasu sformułowania zgadanienia, rola genotypu w kształtowaniu się obserwowalnego fenotypu osobnika jest ciągłym tematem badań [Pigliucci, 2010, Gjuvsland et al., 2013], jednak zrozumienie tej zależności pozostaje wciąż pytaniem otwartym. Prostym przykładem, opisującym jak związek ten jest nieodgadniony, są monozygotyczne (tj. identyczne) bliźnięta, które mają ten sam genotyp, ale nigdy nie mają w pełni tego samego fenotypu, ponieważ chociażby ich odciski palców nigdy nie są całkowicie identyczne. Jak to możliwe? W tej pracy zostały podjęte próby odpowiedzi na pytania dotyczące tego, w jaki sposób poszczególne warstwy wywodzące się z genotypu (transkryptom lub metabolom), wpływają na obserwowalne właściwości fenotypowe, np. formy aktywności w populacjach komórek, typy komórkowe lub podtypów morfologiczne raka.

Wobec powyższego, oprócz teoretycznego aspektu prezentacji nowatorskich algorytmów i metod obliczeniowych, wspólnym celem wszystkich badań naukowych opisanych w dalszej

części rozprawy, z biologicznego punktu widzenia, jest pogłębienie, choćby w niewielkim stopniu, naszego zrozumienia związku genotyp–fenotyp. W kolejnych sekcjach zostaną pokrótce opisane stosowane metody oraz uzyskane dzięki nim wyniki, które stanowią przyczynek do lepszego poznania tego związku.

2 Główne wyniki

Jak już zostało wspomniane, wszystkie wyniki przedstawione w tej rozprawie pochodzą z interdyscyplinarnych badań, co oznacza, że zostały zaproponowane nowe metody obliczeniowe, które są stosowane do pracy nad konkretnymi zagadnieniami biologicznych. To dlatego, każdy z czterech rozdziałów omawia pewien specyficzny problem, który jest następnie rozwiązywany w innowacyjny sposób. W każdym przypadku podany zostaje opis znalezionego rozwiązania i towarzyszące studium przypadku, które zostało przeprowadzone na podstawie rzeczywistych danych biomedycznych. Omawiana rozprawa obejmuje następujące wyniki naukowe.

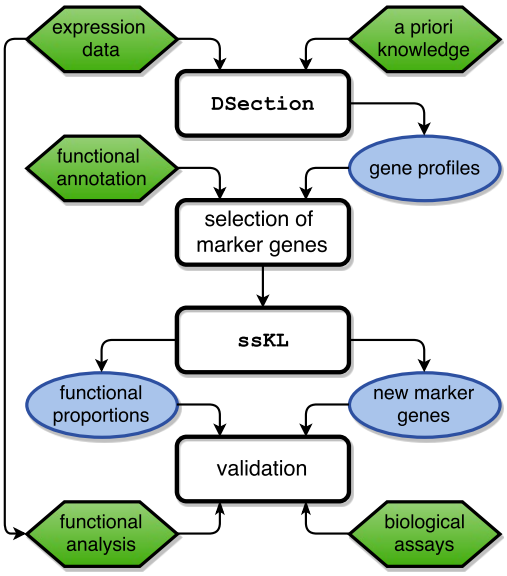
2.1 Dekompozycja sygnału transkryptomicznego

W Rozdziale 2 uwaga zostaje skupiona na problemie dekompozycji uśrednionego sygnału transkryptomicznego pochodzącego z pomiarów aktywności transkrypcyjnej w homogenicznych populacjach komórek. W ogólności, w przypadku większości badań, które koncentrują się na wykrywaniu zmian transkrypcyjnych spowodowanych specyficznymi warunkami eksperymentalnymi, badacze używają próbek biologicznych, składających się z wielu komórek (np. fragmentów tkanki). W takim przypadku, wnioskowanie dotyczące istotnych zmian aktywności genów jest oparte na analizie zachowania stosunkowo dużej populacji komórek, poprzez uśrednianie ich właściwości. Jednakże, nawet przy założeniu całkowitej homogeniczności próbki w kontekście typu komórkowego różne subpopulacje komórek mogą wykazywać odmienne profile transkryptomiczne, ponieważ podlegają one różnymi procesom w ramach szlaków regulatorowych. Celem badania opisanego w Rozdziale 2 jest przedstawienie nowatorskiego schematu metodologicznego uwzględniającego wewnętrzną, funkcjonalną heterogeniczność w jednorodnych liniach komórkowych.

Zaproponowana zostaje metoda obliczeniowa, pozwalająca na wnioskowanie nt. procentowego udziału subpopulacji komórek, które można scharakteryzować aktywnością konkretnych procesów molekularnych zachodzących, w zadanej próbce. Dodatkowo, metoda estymuje profil aktywności transkryptomicznej, dla każdego z wykrytych procesów. W metodzie przyjęto, że obserwowana aktywność transkryptomiczna $A_{i,j}$ określonego genu i w próbce j może być modelowana jako suma aktywności genu i w każdej z k subpopulacji $W_{i,l}$, gdzie $1 \leq l \leq k$, ważona przez współczynnik $H_{l,j}$ opisujący procentowy udział odpowiedniej subpopulacji w całej próbce

$$A_{i,j} = \sum_{l=1}^{k} W_{i,l} H_{l,j} + E_{i,j}$$

gdzie $E_{i,j}$ jest błędem aproksymacji. Opierając się na tym założeniu, wykorzystane zostały dwa algorytmy nieujemnej faktoryzacji macierzy (ang. *Non-negative Matrix Factorization*, NMF) zaproponowane przez [Erkkila et al., 2010] oraz [Brunet et al., 2004]. Rozszerzając je ponadto o dodatkową, biologiczną wiedzą ekspercką zostaje sformułowana metoda `MPH` (ang. *Molecular Process Heterogeneity*), która na podstawie profilu transkryptomicznego badanej populacji komórek, pozwala oszacować: (i) profil transkryptomiczny poszczególnych procesów molekularnych, którym ulegają komórki składające się na całą populację; (ii) procentowy podział aktywności molekularnej w ramach całej populacji komórkowej (patrz Rysunek 2).



Rysunek 2: **Schemat działania metody `MPH`.** Metoda przeprowadza dwie fazy nieujemnej faktoryzacji macierzy, w efekcie dostarczając oszacowanie funkcjonalnego składu homogenicznej populacji komórek wraz z potencjalnymi genami markerowymi i schematami aktywności specyficznych procesów molekularnych.

Ponadto, metoda ta może pomóc w odkryciu potencjalnych biomarkerów charakteryzujących specyficzne procesy molekularne, co stanowi jedno z podstawowych wyzwań współczesnej diagnostyki medycznej.

Poprawność metody `MPH` została zweryfikowana przy użyciu danych transkryptomicznych, pochodzących z homogenicznej linii komórkowej neuroblastomy, wykonanych za pomocą mikromacierzy RNA. Eksperyment miał na celu zbadanie żywotności komórek w dwóch warunkach eksperymentalnych: traktowanie komórek C2–ceramidem oraz C2–ceramidu z towarzyszącym inhibitorem PARP, PJ34. Przedstawiona metodologia jest łatwo adaptowalna do danych transkryptomicznych pochodzących z innych technologii, jak np. RNA-seq lub single-cell RNA-seq. Co więcej, uzupełnia standardowe narzędzia do wskazywania najaktywniejszych ścieżek sygnałowych na podstawie danych transkryptomicznych, a w szczególności może być przydatna w analizie linii komórek rakowych poddanych ekspozycji na biologicznie czynne związki lub leki.

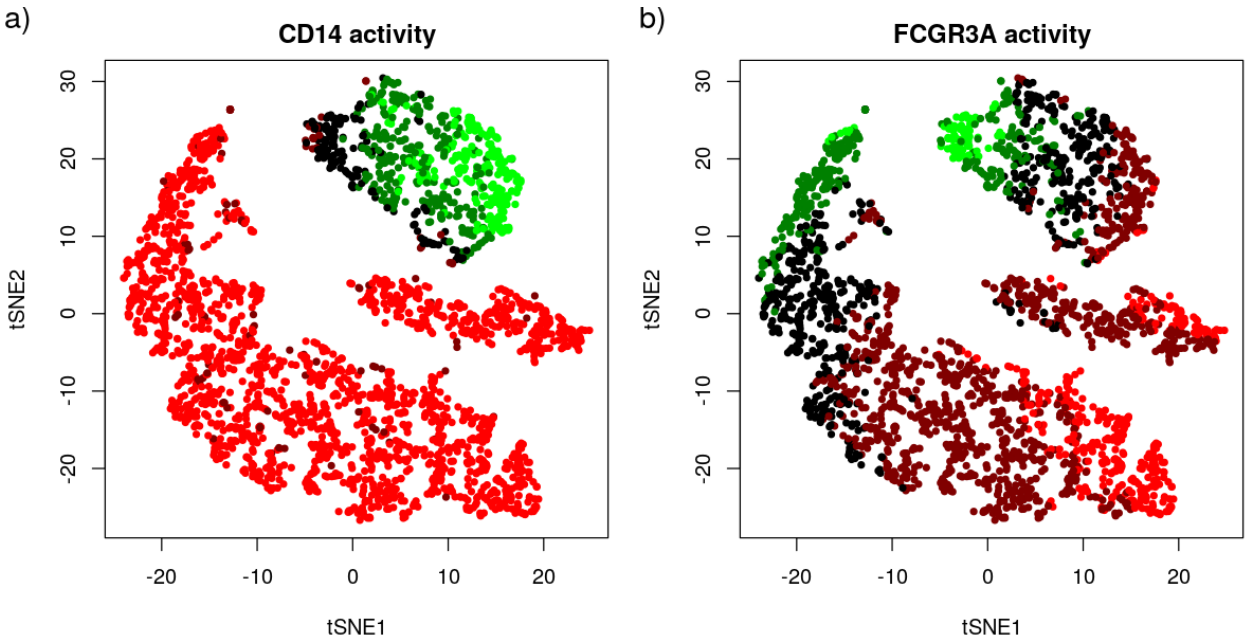
Wyniki przedstawione w tym rozdziale zostały opublikowane wraz z dodatkowym studium przypadku bazującym na analizie danych transkryptomicznych pochodzących z komórek raka jajnika w artykule [Gogolewski et al., 2017].

2.2 Ukryta struktura danych i redukcja szumu

W Rozdziale 3 opisana została metoda ekstrakcji systematycznych cech z danych transkryptomicznych, która nie wykorzystuje zbioru uczącego, czyli tzw. metoda bez nadzoru (ang. *unsupervised*). W literaturze istnieje wiele różnych podejść do rozwiązania tego problemu. Przedstawione w pracy podejście bazuje na istniejącym wariancie algorytmu PCA (`RPCA`) (Robust Principal Component Analysis) zaproponowanym przez [Candès et al., 2011].

Formalnie, algorytm `RPCA` ma na celu dekompozycję macierzy wejściowej A na sumę macierzy niskiego rzędu (ang. *low-rank*) L i macierzy rzadkiej (ang. *sparse*) S poprzez minimalizację następującego problemu optymalizacyjnego:

$$\min_{L,S} ||L||_* + \lambda_1 ||S||_1, \text{ where } A = L + S$$



Rysunek 3: **Poziomy aktywności genów CD14 i FCGR3A.** Panele przedstawiają aktywność genów markerowych wśród wszystkich komórek PBMC użytych w badaniach: (a) CD14, (b) FCGR3A. Algorytm `tRPCAL2` pozwolił na wykrycie systematycznej zmiany aktywności transkryptomicznej tych genów, szczególnie wśród monocytów. Poziom aktywności genów (od najniższej do najwyższej) jest oznaczony skalą koloru od czerwonego, przez czarny do zielonego.

gdzie $||A||_*$ jest normą nuklearną macierzy A, zaś $||A||_1$ jest pierwszą normą wektorowej postaci macierzy A. Początkowo algorytm został zastosowany do analizy sekwencji klatek filmowych i miał na celu wyodrębnienie stałych oraz zmiennych elementów w sekwencji. Jako, że w danych transkryptomicznych również można omawiać wspólne i zmieniające się wzorce, oryginalny algorytm został dostosowany do danych biomedycznych, poprzez opracowanie dwóch rozszerzeń algorytmu `RPCA`.

Po pierwsze, przedstawiona została wersja algorytmu `RPCA`, która ogranicza maksymalny rząd macierzy L. Rozszerzenie bazuje na przyciętej (ang. *truncated*) wersji algorytmu SVD, co oznacza, że w rozkładzie SVD macierzy $A = U\Sigma V^*$ wyznaczona zostaje tylko pożądana liczba (k) wektorów kolumnowych w U i k wektorów wierszowych w V^*, odpowiadających k największym wartościom singularnym z macierzy Σ. Rozszerzenie to, nazwane `tRPCA` (truncated RPCA), sprawia, że algorytm jest szybszy i zużywa mniej pamięci operacyjnej niż pierwotna wersja `RPCA`. Dodatkowo, ponieważ ta wersja algorytmu nakłada dodatkowe ograniczenie na maksymalny rząd macierzy L powstałej podczas dekompozycji, umożliwia to określenie oczekiwanego wymiaru opisującego niezmienną część danych.

Następnie, ponieważ dane transkryptomiczne z natury zawierają biologiczne fluktuacje różnego pochodzenia, przypuszczalnie, stała część macierzy niskiego rzędu L będzie przechowywała niechciane efekty stochastyczne. Z tego powodu wprowadzone zostało kolejne rozszerzenie algorytmu `tRPCA` dokonujące dodatkowej redukcji gęstego szumu za pomocą regularyzacji L_2 (`tRPCAL2`) (patrz Alg. 1). W przeciwieństwie do algorytmów `RPCA` i `tRPCA`, które uwzględniają w dekompozycji jedynie macierz niskiego rzędu L i rzadką S, proponowany algorytm wyodrębnia dodatkowo macierz szumu E. W takim ujęciu, początkowy problem

Algorithm 1 truncated-RPCA with L2 regularization

procedure tRPCAL2(λ_1, λ_2, k_0, c)
  $S_0, Y_0, E_0 \leftarrow 0$; $\mu_0 > 0$; $k = k_0$
  while not converged **do**
    compute $L_{i+1} = \mathcal{D}_{\mu_i^{-1}}(A - S_i - E_i + \mu_i^{-1} Y_i)$
    compute $E^*_{i+1} = \mathcal{E}_{\lambda_2 \mu_i^{-1}}(A - S_i - L_{i+1} + \mu_i^{-1} Y_i)$
    compute $S_{i+1} = \mathcal{S}_{\lambda_1 \mu_i^{-1}}(A - E^*_{i+1} - L_{i+1} + \mu_i^{-1} Y_i)$
    compute $E_{i+1} = \mathcal{E}_{\lambda_2 \mu_i^{-1}}(A - S_{i+1} - L_{i+1} + \mu_i^{-1} Y_i)$
    compute $Y_{i+1} = Y_i + \mu_i \cdot (A - E_{i+1} - L_{i+1} - S_{i+1})$
    compute $\mu_{i+1}^{-1} = \max(c \cdot \mu_i^{-1}, \sigma_{k+1})$
    if $\mu_i^{-1} == \sigma_{k+1}$ **then** increase k
    else $k = 1 + \text{argmax}_j\,(\sigma_j > \mu_{i+1}^{-1})$
    end if
  end while
end procedure

optymalizacyjny został rozszerzony do następującej postaci:

$$A = L + S + E$$
$$\min_{L,S,E} ||L||_* + \lambda_1 ||S||_1 + \lambda_2 ||E||_F$$

gdzie $||A||_F$ jest normą Frobeniusa macierzy A. W celu rozwiązania tego zagadnienia, metoda *Alternating Direction* została rozszerzona i wprowadzono dodatkowy operator odcięcia (ang. *threshold operator*)

$$\mathcal{E}_\tau(X) = \max\left(0, 1 - \frac{\tau}{||X||_F}\right) \cdot X$$

który określa jakie wartości mogą występować w macierzy E.

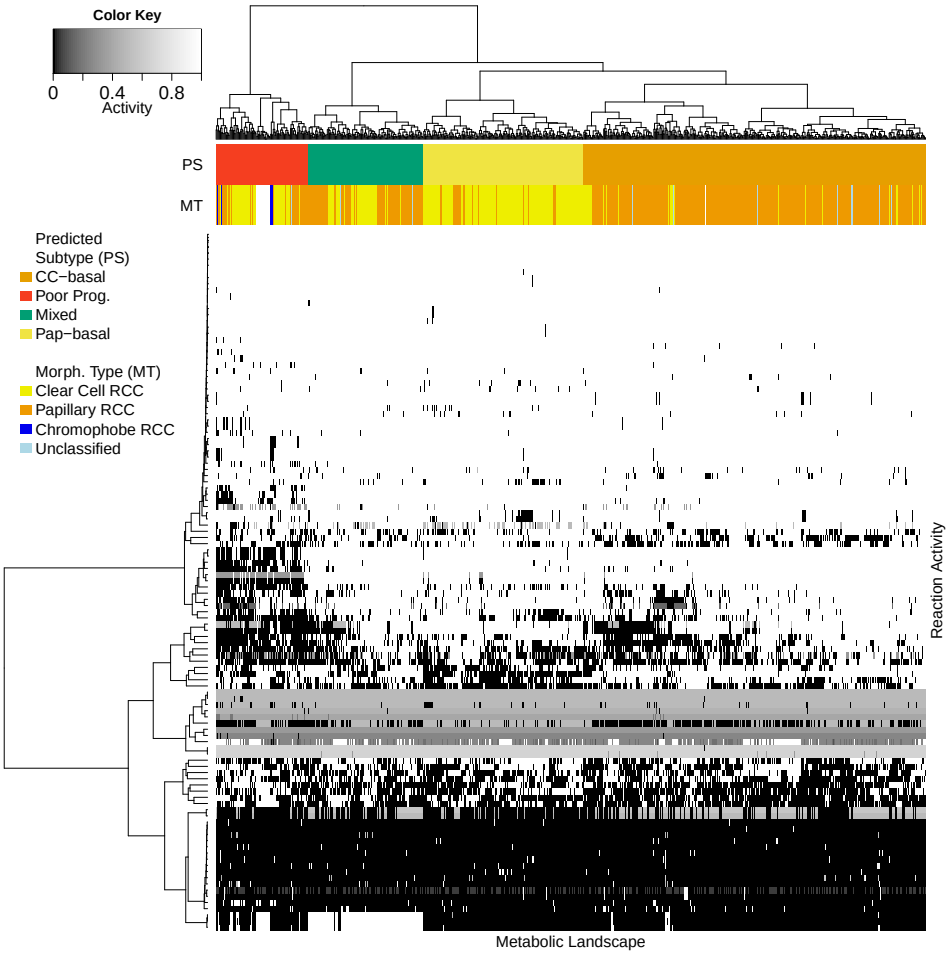
W ostatniej części Rozdziału 3 omówione zostaje zastosowanie zaproponowanej metody do danych single-cell RNA-seq pochodzących z jednojądrzastych komórek krwi obwodowej (ang. *peripheral blood mononuclear cells*, PBMC). W szczególności, opisano w jaki sposób, wykorzystując macierz niskiego rzędu L, można osiągnąć lepszą klasyfikację nieoznaczonych pojedynczych komórek. Co więcej, zostało zaprezentowane zastosowania algorytmu `tRPCAL2` do wykrywania koekspresji genów oraz detekcji podtypów komórkowych (patrz Rysunek 3). Na zakończenie podkreślona została możliwości zastosowania proponowanych wariantów algorytmów do innych zaszumionych danych wielkoskalowych.

Większa część wyników przedstawionych w tym rozdziale została już opublikowana w materiałach pokonferencyjnych *14th International Symposium on Bioinformatics Research and Applications* (ISBRA) [Gogolewski et al., 2018b]. Dodatkowo, po otrzymaniu zaproszenia do publikacji rozszerzonej wersji artykułu, jest ona aktualnie na etapie recenzji w czasopiśmie *Journal of Computational Biology*.

2.3 Problem nadreprezentacji w modelach sieci metabolicznych

Ogromna ilość dostępnych danych molekularnych pozyskiwanych z wielu różnych źródeł otwiera nowe możliwości integracji danych. Niemniej jednak, proces integracji powinien być prowadzony z zachowaniem ostrożności w kontekście szeroko rozumianego problemu przeuczania modeli statystycznych. W Rozdziale 4 przedstawione zostały dwa rodzaje wyników.

Z jednej strony zaprezentowano, w jaki sposób integracja wiedzy metabolomicznej i transkryptomicznej może doprowadzić do redundancji statystycznej, co skutkuje wnioskowaniem



Rysunek 4: **Grupowanie względem aktywności reakcji metabolicznych.** Rysunek przedstawia jakość grupowania aktywności reakcji próbek raka nerki po zastosowaniu korekty struktury sieci metabolicznej. Dwa poziome, kolorowe pasy opisują przyporządkowanie próbek do dwóch grup: przewidywanych podtypów (PS) oraz znanych typów morfologicznych (MT).

na temat artefaktów związanych ze sposobem integracji. Ponadto, pokazano, że wnioski te mogą być poparte literaturą naukową, która sugerowałaby ich poprawność. Wynik ten ma wartość meta-naukową, ponieważ pokazuje, jak doniesienia naukowe mogą wprowadzać nierównowagę wiedzy w modelach matematycznych opisujących zjawiska biologiczne. Celem lepszego objaśnienia tego problemu, przedstawiony został przykład integracji danych RNA-seq z modelem metabolizmu ludzkiego RECON 2.2 [Swainston et al., 2016] i opisane wyniki tego badania. Uzyskane rezultaty sugerują istnienie biomarkerów metabolomicznych, które mogą wyraźnie rozróżniać grupy rakowych pacjentów poprzez określony wzór aktywności metabolicznej. Co więcej, opisane odkrycia zostały poparte współczesnymi wynikami badań zawartymi w literaturze. Jednakże, jako kontrprzykład dla początkowych odkryć, została również przedstawiona analiza przeprowadzona na losowym zbiorze danych. Uzyskany wynik dowiódł, że wszystkie obserwacje należy traktować jako artefakty wynikające z nieregularności informacji użytej do konstrukcji używanego modelu.

Wobec powyższego, aby rozwiązać ten problem, w dalszej części Rozdziału 4 została zasugerowana możliwa metoda redukcji obciążenia (ang. *bias*), która poddana została dodatkowej weryfikacji za pomocą zestawu danych transkryptomicznych z bazy TCGA (The Cancer Genome Atlas) opisujących pacjentów ze zdiagnozowanym rakiem nerki (ang. *Renal Cell Carcinoma*, RCC). Dzięki redukcji, został opracowany wiarygodny, metaboliczny

opis znanych morfologicznych podtypów RCC. Ponadto, wyodrębniona została grupa słabo rokujących pacjentów, charakteryzująca się niską aktywnością enzymów odpowiadających za transport leków istotnych dla procesu chemioterapii.

Wyniki przedstawione w Rozdziale 4 mają stanowić znak ostrzegawczy, który przypomina innym badaczom, że modele integracyjne powinny być starannie przygotowane zarówno pod względem formalnym, tj. matematycznym, jak i praktycznym, tj. zgodnym z rzeczywistością. W przeciwnym razie nie powinno dziwić, że jeśli założenia modelu opierają się na nierównomiernie dobranych raportach naukowych, to ich wyniki mogą być poparte tymi samymi źródłami.

Główne wyniki opisane w tym rozdziale zostały zaprezentowane na 2018 *IEEE International Conference on Bioinformatics and Biomedicine* (BIBM) i opublikowane w materiałach konferencyjnych [Gogolewski et al., 2018a]. Dodatkowo, studium przypadku dla danych single-cell RNA-seq zaprezentowane zostało w ramach *Workshop on Computational Advances for Single-Cell Omics Data Analysis* podczas *2018 IEEE International Conference on Computational Advances in Bio and medical Sciences* (ICCABS) [Gogolewski and Gambin, 2018].

2.4 Interdyscyplinarne badania biomedyczne

W Rozdziale 5 zostały omówione wyniki analiz bioinformatycznych, koncentrujących się na wyborze cech w kontekście różnego rodzaju danych eksperymentalnych mających na celu zrozumienie roli genu FOXF1 w rozwoju płuc. W szczególności, opisano wyniki analizy dotyczącej wpływu zmiany liczby kopii genu FOXF1 na regulację jego aktywności transkrypcyjnej. Ponadto przedstawiona została analiza porównawcza rozmieszczenia sekwencji minisatelit w różnych gatunkach i omówiono ich potencjalną rolę, jako źródła błędów replikacji DNA.

Cała praca z Rozdziału 5 jest wynikiem interdyscyplinarnej współpracy z Baylor Collage of Medicine i stanowi część już opublikowanych artykułów: [Dharmadhikari et al., 2014, Dharmadhikari et al., 2016].

3 Współpraca naukowa

W prowadzone badania, zaprezentowane w ramach tej rozprawy, zaangażowani byli różni współpracownicy oraz grupy naukowe. Grupa prof. Strosznjadera z Instytutu Medycyny Doświadczalnej i Klinicznej im. M. Mossakowskiego, PAN przeprowadziła wszystkie eksperymenty laboratoryjne i dostarczyła dane mikromacierzowe dla studium przypadku przedstawionego w Rozdziale 2. Podobnie grupa prof. Lesyng uczestniczyła w biologicznej interpretacji wyników analiz obliczeniowych. Grupa prof. Stankiewicza z Baylor Collage of Medicine koordynowała badania, na podstawie których przeprowadzono wszystkie analizy bioinformatyczne przedstawione w Rozdziale 5. Wreszcie, grupa doktora Le Rouzic'a, współtworzyła projekt poboczny związany z ewolucją ruchomych elementów genetycznych.

Ponadto, w przeprowadzanie badań przedstawionych w tej pracy i opublikowanych w ramach kilku artykułów naukowych zaangażowani byli także inni współpracownicy naukowi. W [Gogolewski et al., 2017] badania wspierali biolodzy: W. Wronowska uczestniczyła w biologicznej interpretacji i opisie wszystkich wyników oraz oceniała biologiczną poprawność modelu; A. Lech, zinterpretowała wyniki drugiego studium przypadku zaprezentowanego w artykule. M. Sykulski pomagał w projektowaniu i wdrażaniu algorytmów w [Gogolewski et al., 2018b], zaś N. C. Chung zaproponował wykorzystanie danych scRNA-seq. Dodatkowo, M. Kostecki ponownie zaimplementował algorytm użyty w [Gogolewski et al., 2018a] i pomógł w analizie modelu RECON 2.2. Wreszcie, A. Gambin inspirowała i koordynowała całą pracę prezentowaną w tej rozprawie oraz opublikowanych artykułach.

Wykaz publikacji zawierających główne wyniki przedstawione w rozprawie

Gogolewski K., Wronowska W., Lech A., Lesyng B., Gambin A. (2017) Inferring Molecular Processes Heterogeneity from Transcriptional Data. *Biomedical Research International*, 2017:6961786.

Gogolewski K., Kostecki M., Gambin A. (2018a) Renal cell carcinoma classification: a case study of pitfalls associated with metabolic landscape analysis. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 96–101.

Gogolewski K., Sykulski M., Chung N.C., Gambin A. (2018b) Truncated Robust Principal Component Analysis and Noise Reduction for Single Cell RNA-seq Data. In: Zhang F., Cai Z., Skums P., Zhang S. (eds) *Bioinformatics Research and Applications. ISBRA 2018. Lecture Notes in Computer Science*, vol 10847. Springer, Cham.

Wykaz innych publikacji

Dharmadhikari A.V., Gambin T., Szafranski P., Cao W., Probst F.J., Jin W., Fang P., Gogolewski K., Gambin A., George-Abraham J.K., Golla S., Boidein F., Duban-Bedu B., Delobel B., Andrieux J., Becker K., Holinski-Feder E., Cheung S.W., Stankiewicz P. (2014) Molecular and clinical analyses of 16q24.1 duplications involving FOXF1 identify an evolutionarily unstable large minisatellite. *BMC Medical Genetics*, 15:128.

Gogolewski K., Startek M., Gambin A., Le Rouzic A. (2016) Modelling the proliferation of transposable elements in populations under environmental stress. *ArXiv e-prints, Quantitative Biology - Populations and Evolution*, 92-08, J.3, arXiv:1611.04812.

Dharmadhikari A.V., Sun J.J., Gogolewski K., Carofino B.L., Ustiyan V., Hill M., Majewski T., Szafranski P., Justice M.J., Ray R.S., Dickinson M.E., Kalinichenko V.V., Gambin A., Stankiewicz P. (2016) Lethal lung hypoplasia and vascular defects in mice with conditional Foxf1 overexpression. *Biology Open*, 5, 11:1595-1606.

Gogolewski K., Gambin A. (2018) PCA-like Methods for the Integration of Single Cell RNA-seq Data with Metabolic Networks. In: 2018 *IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*. Springer.

References

[Alberch, 1991] Alberch, P. (1991). From genes to phenotype: dynamical systems and evolvability. *Genetica*, 84(1):5–11.

[Borgan et al., 2010] Borgan, E., Sitter, B., Lingjaerde, O. C., Johnsen, H., Lundgren, S., Bathen, T. F., Sorlie, T., Borresen-Dale, A. L., and Gribbestad, I. S. (2010). Merging transcriptomics and metabolomics–advances in breast cancer profiling. *BMC Cancer*, 10:628.

[Borrageiro et al., 2018] Borrageiro, G., Haylett, W., Seedat, S., Kuivaniemi, H., and Bardien, S. (2018). A review of genome-wide transcriptomics studies in Parkinson's disease. *Eur. J. Neurosci.*, 47(1):1–16.

[Brunet et al., 2004] Brunet, J. P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.*, 101(12):4164–4169.

[Candès et al., 2011] Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37.

[Dharmadhikari et al., 2014] Dharmadhikari, A. V., Gambin, T., Szafranski, P., Cao, W., Probst, F. J., Jin, W., Fang, P., Gogolewski, K., Gambin, A., George-Abraham, J. K., Golla, S., Boidein, F., Duban-Bedu, B., Delobel, B., Andrieux, J., Becker, K., Holinski-Feder, E., Cheung, S. W., and Stankiewicz, P. (2014). Molecular and clinical analyses of 16q24.1 duplications involving FOXF1 identify an evolutionarily unstable large minisatellite. *BMC Med. Genet.*, 15:128.

[Dharmadhikari et al., 2016] Dharmadhikari, A. V., Sun, J. J., Gogolewski, K., Carofino, B. L., Ustiyan, V., Hill, M., Majewski, T., Szafranski, P., Justice, M. J., Ray, R. S., Dickinson, M. E., Kalinichenko, V. V., Gambin, A., and Stankiewicz, P. (2016). Lethal lung hypoplasia and vascular defects in mice with conditional Foxf1 overexpression. *Biol Open*, 5(11):1595–1606.

[Erkkila et al., 2010] Erkkila, T., Lehmusvaara, S., Ruusuvuori, P., Visakorpi, T., Shmulevich, I., and Lahdesmaki, H. (2010). Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, 26(20):2571–2577.

[Gjuvsland et al., 2013] Gjuvsland, A. B., Vik, J. O., Beard, D. A., Hunter, P. J., and Omholt, S. W. (2013). Bridging the genotype-phenotype gap: what does it take? *J. Physiol. (Lond.)*, 591(8):2055–2066.

[Gogolewski and Gambin, 2018] Gogolewski, K. and Gambin, A. (2018). Pca-like methods for the integration of single cell rna-seq data with metabolic networks. In *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, pages 1–1.

[Gogolewski et al., 2018a] Gogolewski, K., Kostecki, M., and Gambin, A. (2018a). Renal cell carcinoma classification: a case study of pitfalls associated with metabolic landscape analysis. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 96–101.

[Gogolewski et al., 2018b] Gogolewski, K., Sykulski, M., Chung, N. C., and Gambin, A. (2018b). Truncated robust principal component analysis and noise reduction for single

cell rna-seq data. In Zhang, F., Cai, Z., Skums, P., and Zhang, S., editors, *Bioinformatics Research and Applications*, pages 335–346, Cham. Springer International Publishing.

[Gogolewski et al., 2017] Gogolewski, K., Wronowska, W., Lech, A., Lesyng, B., and Gambin, A. (2017). Inferring Molecular Processes Heterogeneity from Transcriptional Data. *Biomed Res Int*, 2017.

[Hasin et al., 2017] Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.*, 18(1):83.

[Li et al., 2016] Li, S., Todor, A., and Luo, R. (2016). Blood transcriptomics and metabolomics for personalized medicine. *Comput Struct Biotechnol J*, 14:1–7.

[Lowe et al., 2017] Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLoS Comput. Biol.*, 13(5):e1005457.

[Pigliucci, 2010] Pigliucci, M. (2010). Genotype-phenotype mapping and the end of the 'genes as blueprint' metaphor. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 365(1540):557–566.

[Ren et al., 2016] Ren, S., Shao, Y., Zhao, X., Hong, C. S., Wang, F., Lu, X., Li, J., Ye, G., Yan, M., Zhuang, Z., Xu, C., Xu, G., and Sun, Y. (2016). Integration of Metabolomics and Transcriptomics Reveals Major Metabolic Pathways and Potential Biomarker Involved in Prostate Cancer. *Mol. Cell Proteomics*, 15(1):154–163.

[Reuter et al., 2015] Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Mol. Cell*, 58(4):586–597.

[Riekeberg and Powers, 2017] Riekeberg, E. and Powers, R. (2017). New frontiers in metabolomics: from measurement to insight. *F1000Res*, 6:1148.

[Shin et al., 2018] Shin, T. H., Lee, D. Y., Lee, H. S., Park, H. J., Jin, M. S., Paik, M. J., Manavalan, B., Mo, J. S., and Lee, G. (2018). Integration of metabolomics and transcriptomics in nanotoxicity studies. *BMB Rep*, 51(1):14–20.

[Stempler et al., 2014] Stempler, S., Yizhak, K., and Ruppin, E. (2014). Integrating transcriptomics with metabolic modeling predicts biomarkers and drug targets for Alzheimer's disease. *PLoS ONE*, 9(8):e105383.

[Swainston et al., 2016] Swainston, N., Smallbone, K., Hefzi, H., and Dobson et al., P. D. (2016). Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, 12:109.