Analiza różnicowej ekspresji genów w populacjach bakterii

Autoreferat

Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski

Julia Herman-Iżycka

maj 2021

1 Wstęp

Bakterie oraz inne mikroorganizmy można spotkać niemalże w każdym środowisku. Zamieszkują nie tylko glebę czy oceany, ale też ludzkie ciało. Poszczególne regiony ciała, takie jak skóra, jama ustna, czy jelita, zapewniają różne warunki środowiskowe, więc są zamieszkiwane przez odrębne, wielogatunkowe populacje mikroorganizmów – mikrobioty.

Bakterie jelitowe stanowią większą część mikroorganizmów zamieszkujących ciało człowieka. Szacuje się, że u jednej osoby w jelitach żyje ok. 1000 gatunków bakterii, a ponadto pewna liczba archeonów, grzybów i protistów [Lozupone et al., 2012]. Skład mikrobioty zależy od wielu czynników, w tym diety [David et al., 2014], genetyki gospodarza [Goodrich et al., 2014], wieku czy miejsca zamieszkania. Duże zainteresowanie badaniami mikrobioty jelitowej wynika nie tylko z jej liczności ale również z faktu, że ze względu na zamieszkiwanie jelit mikroorganizmy te mają wpływ nie tylko na trawienie i wchłanianie pewnych substancji odżywczych, np. krótkołańcuchowych kwasów tłuszczowych i witamin, ale też funkcjonowanie układu odpornościowego, a mogą one nawet wpływać na układ nerwowy. Relacja mikrobiom-jelita-mózg jest nazywana także osią jelita-mózg i polega na dwukierunkowej biochemicznej komunikacji między przewodem pokarmowym a nerwowym. Odbywa się ona prawdopodobnie za pośrednictwem neuromodulatorów oraz nerwu błędnego [Sgritta et al., 2019]. Do tej pory wykazano wpływ mikrobioty np. na stany lękowe i depresję [Foster and Neufeld, 2013].

Dzięki niedawnemu rozwojowi technologii sekwencjonowania wysokoprzepustowego możliwe staje się badanie nie tylko genomów pojedynczych szczepów bakterii, które udało się wyhodować poza środowiskiem ich życia, ale także genomów całych populacji mikroorganizmów – mikrobiomów. Technologie sekwencjonowania nowej generacji takie jak Illumina, czy trzeciej generacji – np. Nanopore, pozwalają uzyskać znaczne liczby odczytów reprezentujących fragmenty sekwencji DNA lub RNA, w zależności od eksperymentu. Ponieważ jednak odczyty są wciąż krótsze niż genomy bakteryjne (w przypadku nadal popularnej drugiej generacji są to setki par zasad), a w odczytach zdarzają się błędy, to identyfikacja na ich podstawie zawartości próbki prowadzi do ciekawych problemów natury obliczeniowej.

1.1 Wstęp do problemów dotyczących analizy danych metatranskryptomicznych

W zrozumieniu mechanizmu wpływu mikrobioty jelitowej na człowieka, ale też innych mechanizmów oddziaływania mikroorganizmów na ich środowisko życia, pomocne może być zbadanie aktywności genów. W takich badaniach sekwencjonowane jest RNA (po usunięciu rybosomalnego RNA) aby oszacować ilość mRNA, które wskazuje które geny i jak często są aktywowane. Często porównywany jest również wpływ na ekspresję genów danego czynnika poprzez zebranie próbek ze środowiska poddanego wpływowi badanego czynnika oraz bez jego wpływu.

W badaniach metatranskryptomicznych pojawiają się zatem dwa rodzaje wyzwań – identyfikacja obecnych w próbce bakterii i ich aktywnych genów, oraz ilościowe badanie ekspresji genów i związana z nim identyfikacja genów zmieniających się pod wpływem działania czynnika. Formalny opis dwóch podejść pojawiających się w analizach danych metatranskryptomicznych zawarty jest w rozdziale drugim. W pierwszej jego części szeroko opisany został problem składania sekwencji na podstawie krótkich jej fragmentów, czyli asemblacji *de novo*, zarówno przy pomocy grafu nałożeń [Myers, 1995] jak i aktualnie częściej używanego grafu *de Bruijna* [Pevzner et al., 2001]. Omówione zostały częste wyzwania pojawiające się w asemblacji danych z sekwencjonowania wysokoprzepustowego związane z technologią (błędy) i z naturą sekwencji biologicznych (np. polimorfizm sekwencji, powtórzenia), a które objawiają się w obu rodzajach grafów jako pętle, bąble, odstające końcówki czy splątania.

Kolejna część rozdziału drugiego traktuje o problemie mapowania odczytów na sekwencje referencyjne i formalizuje kryteria zliczania odczytów dopasowanych do sekwencji referencyjnych, czyli obliczania pokrycia. Wprowadza też definicję krotności zmiany (ang. *fold change*). Rozdział drugi zawiera także omówienie wybranych programów z szerokiego zakresu istniejących implementacji sposobów rozwiązania problemów: asemblacji *de novo*, mapowania i analizy różnicowej ekspresji. Omówione zostały programy, których użyteczność do danych metatranskryptomicznych jest przedmiotem dalszej części pracy, m.in. programy do asemblacji `Velvet` [Zerbino and Birney, 2008], `MEGAHIT` [Li et al., 2015], `SGA` [Simpson and Durbin, 2012], programy do mapowania `bowtie2` [Langmead and Salzberg, 2012], `minimap2` [Li, 2018] oraz `kallisto` [Bray et al., 2016], program do analizy danych transkryptomicznych poprzez pseudomapowanie, którego użyteczność dla danych metagenomicznych pokazano wcześniej w pracy [Schaeffer et al., 2017].

2 Najważniejsze wyniki

Zaprezentowane w rozprawie metody analizy danych metatranskryptomicznych zostały przetestowane na danych eksperymentalnych pochodzących z sekwencjonowania mRNA uzyskanego z mysich odchodów. Próbki pochodziły z eksperymentu, w którym 5 myszy – 3 myszy typu dzikiego i 2 myszy trisomiczne (których genotyp jest mysim odpowiednikiem zespołu Downa u ludzi), zostały poddane zmianie diety z normalnej na dietę wysokotłuszczową. Próbki pochodziły z dwóch punktów czasowych – krótko po zmianie diety i po 2 tygodniach diety wysokotłuszczowej. Uzyskano w ten sposób 9 próbek, które zsekwencjonowano otrzymując zbiory odczytów zawierające średnio kilka milionów par odczytów o długościach 100 bp każdy i dobrej jakości sekwencjonowania.

2.1 Analiza danych metatranskryptomicznych poprzez mapowanie

W rozdziale 3 omówione zostało zastosowanie standardowych metod obliczeniowych stosowanych w analizie danych metagenomicznych i transkryptomicznych, gdyż metod *stricte* metatranskryptomicznych wciąż brakuje. W pierwszej części rozdziału trzeciego przedstawione jest mapowanie na różne zbiory sekwencji referencyjnych przy użyciu znanych metod mapowania: `bowtie2` i `minimap2`, oraz metody do szacowania liczności transkryptów na podstawie odczytów bez mapowania – `kallisto`. Porównane zostały różne zbiory sekwencji referencyjnych, m.in. genomy bakteryjne z bazy RefSeq [Pruitt et al., 2007], genomy bakterii jelitowych zebrane przez *Human Microbiome Project* [Peterson et al., 2009] oraz część genomów RefSeq zidentyfikowana przez program Kraken [Wood and Salzberg, 2014].

Z przedstawionych w tabeli 2.1 wyników mapowania można wywnioskować, że dobór narzędzia oraz parametrów mapowania jest bardzo istotny oraz że najlepszym narzędziem do mapowania na różne zbiory genomów referencyjnych jest `kallisto` z wielkością k-meru k = 21, jednak nawet wtedy niewiele ponad 21% odczytów udaje się zmapować jednoznacznie na sekwencje referencyjne, a jednocześnie część odczytów nie jest mapowana z powodu konfliktowego przypisania. Oznacza to, że zwiększanie dostępnej mocy obliczeniowej oraz zwiększanie liczby zsekwencjonowanych genomów prawdopodobnie nie jest rozwiązaniem problemu niskiej mapowalności na bazy sekwencji referencyjnych w przypadku krótkich odczytów.

Zbiór sekwencji	Mapowanie				
	metoda	zmapowane	jednoznacznie	brak k-merów	konflikty
RefSeq2014	k=13	0.19%	0.09%	0	99.81%
	k=21	26.86%	21.06%	32.86%	40.28%
	k=25	11.05%	8.37%	78.12%	10.83%
	k=31	6.6%	4.67%	88.07%	5.32%
	bowtie2	3.64%	0.5%		
	minimap2	8.64%	8.38%		
RefSeq2018	k=21		przekroczone 800GB RAM		
	k=31	12.56%	10.83%	81.05%	6.39%
	bowtie2	9.14%	3.15%		
	minimap2	15.4%	8.58%		
HMP	k=21	23.52%	15.68%	54.77%	21.71%
	bowtie2	4.97%			
miBC	k=21	21.02%	18.37%	68.31%	10.67%
	bowtie2	8.96%			
	minimap2	11.81%	5.98%		
kraken ($\Sigma > 18$)	k=31	12.56%	10.87%	81.06%	6.37%
	bowtie2	9.14%			

Tabela 2.1: Wyniki mapowania na różne zbiory genomów referencyjnych.

2.2 Analiza danych metatranskryptomicznych poprzez asemblację

W dalszej części rozdziału 3 porównane zostały metody asemblacji o różnych docelowych zastosowaniach oraz stosujące różne podejścia. Ponieważ w asemblacji danych metatranskryptomicznych nie są znane docelowe liczby i długości sekwencji, które powinny być odtworzone w asemblacji, za kryterium porównawcze uznana została średnia mapowalność (przy pomocy `kallisto`) odczytów na sekwencje zło-

żone ze zbioru odczytów ze wszystkich próbek, gdyż mapowanie na zasemblowane sekwencje jest jednym ze stosowanych w tej dziedzinie podejść.

Spośród wielu porównanych metod zarówno opracowywanych pod kątem asemblacji pojedynczego genomu (`Velvet`, `SGA`), transkryptomu (`oases`) czy metagenomu (`Metavelvet`, `MEGAHIT`, `IDBA-UD`), najlepsze wyniki udało się osiągnąć dzięki programowi `MEGAHIT`, łączącemu wykorzystanie wielu wartości parametru k – rozmiaru k-meru w grafie, i opartemu na skompresowanych (zwięzłych) grafach *de Bruijna*. Do kontigów zbudowanych przy pomocy tego programu mapowało się ponownie prawie 84% odczytów, a ich liczba i sumaryczna długość nie odstawały od pozostałych metod, choć nie były największe (patrz tabela 2.2).

Program	Liczba kontigów	Sumaryczna długość	zmapowane (jednoznacznie)
Velvet	215,813	144,650,085	49.5% (48.9%)
Oases	635,425	591,707,639	52.8% (16.9%)
Metavelvet	471,289	217,391,825	42.5% (42.0%)
MEGAHIT	237,151	230,704,864	83.9% (83.1%)
IDBA-UD	219,352	228,239,757	77.5% (77.2%)
SGA	232,576	134,778,668	38.1% (37.3%)

Tabela 2.2: Wyniki asemblacji przy pomocy różnych dostępnych programów.

Ponieważ mapowalność była najważniejszym kryterium to właśnie sekwencje z `MEGAHIT` zostały wybrane do dalszych analiz, w tym próby analizy różnicowej i do analizy funkcjonalnej.

2.3 Podejście funkcjonalne

W sekcji 3.5 przedstawione zostały wyniki funkcjonalnej analizy omawianych próbek, wykonane wspólnie z dr Iloną Grabowicz. Zaproponowana procedura obejmowała wybór genów z kontigów z `MEGAHIT` przy pomocy programu `MetaGeneMark` [Zhu et al., 2010] ze względu na dużą, większą niż spodziewana dla genów bakteryjnych długość kontigów (co może sugerować nieprawidłowe „sklejenie" niektórych genów), a także możliwość pojawiania się w przypadku bakterii kilku genów z jednego operonu w jednym transkrypcie. Kolejnym krokiem procedury było mapowanie odczytów na znalezione sekwencje genów. Sekwencjom genów następnie przypisano funkcje (przy pomocy programu `eggNOGmapper` [Huerta-Cepas et al., 2017]) i zsumowano liczby odczytów zmapowanych na geny o tej samej przypisanej funkcji a pochodzące z różnych kontigów. Takie podejście pozwoliło zbadać jak kolektywnie zmieniła się aktywność bakterii gatunkowego składu mikrobioty.

Dopiero dla tak zagregowanych danych udało się uzyskać statystycznie istotnie zmienione geny. Istotnej zmianie na diecie wysokotłuszczowej uległy 4 geny, wzrosła ekspresja genów: *ltaS*, kodującego syntazę kwasu lipotejchojowego, będącego składnikiem ścian komórkowych, i *dhbF*, syntazy peptydów ze ścieżki aktywowanej w warunkach niedoboru żelaza, a spadła ekspresja *proWX* związanego z warunkami wysokiego stresu osmotycznego. Wśród 100 najbardziej zmienionych genów były też geny związane z wirulencją, transportem membranowym czy metabolizmem cukrów i aminokwasów.

Pomiędzy próbkami zebranymi od myszy różniących się genetycznie 137 genów było istotnie zmienionych. U myszy trisomicznych bardziej aktywne były geny me-

tabolizmu hipoksantyny, a u myszy typu dzikiego geny odpowiadające za wirulencję i interakcje z komórkami gospodarza, jak również geny kodujące transpozazy.

Chociaż przy użyciu przedstawionej procedury udało się zidentyfikować różnicowo eksprymowane geny, a zmiany w grupach funkcjonalnych do których one należały są spodziewane, to liczba etapów, które były do tego wymagane, a które są częściowo redundantne, posłużyła jako motywacja do sformułowania i rozważenia opisanego w dalszej części pracy problemu uzyskania różnicowych sekwencji bezpośrednio w wyniku asemblacji.

2.4 Teoretyczna możliwość zastosowania grafu nałożeń i grafu *de Bruijna* w asemblacji różnicowych kontigów

W rozdziale 4 zdefiniowany został formalnie problem asemblacji różnicowych kontigów oraz rozważone zostały możliwości wykorzystania dwóch popularnych metod asemblacji.

W ogólnej definicji problemu różnicowych kontigów użyto krotności zmiany (ang. *fold-change*, ozn. fc) do wyróżnienia kiedy kontig uznany jest za różnicowy.

Definicja 2.1 (Problem asemblacji *FC*-różnicowych kontigów)
Dane są dwa zbiory próbek: traktowane $\mathcal{T} = \{S_1^t, \ldots, S_{n_t}^t\}$ *(ang.* treated*) oraz kontrolne* $\mathcal{C} = \{S_1^c, \ldots, S_{n_c}^c\}$ *(ang.* control*), oraz* FC *– wartość oczekiwanej krotności zmiany.*

Znajdź zbiór sekwencji $\mathcal{R} = \{r_1, r_2, \ldots r_{|\mathcal{R}|}\}$ *oraz mapowanie* $\mathcal{M}_\mathcal{R}$ *takie, że*

- *każda sekwencja* $r_i \in \mathcal{R}$ *jest* FC*-różnicowa:*

$$fc(i) = \frac{\overline{c}_{\mathcal{M}_\mathcal{R},\mathcal{T}}(r_i)}{\overline{c}_{\mathcal{M}_\mathcal{R},\mathcal{C}}(r_i)} \geq FC \quad lub \quad fc(i) \leq \frac{1}{FC} \quad lub \quad \overline{c}_{\mathcal{M}_\mathcal{R},\mathcal{C}}(r_i) = 0$$

  gdzie zliczenia dla zbioru próbek zdefiniujemy jako znormalizowaną sumę zliczeń z próbek $\overline{c}_{m_R,\mathcal{T}}(g_i) = \frac{\sum_j c_{m_R,S_j^t}(g_i)}{\sum_j |S_j^t|}$

- $\mathcal{R}$ *jest wynikiem asemblacji pewnego podzbioru sekwencji* S*, ozn.* $\mathcal{A}$*, czyli*

$$\mathcal{A} = \{s_i \in \mathcal{C} \cup \mathcal{T}\} : \exists_{1 \leq j \leq |G|} s_i \sqsubseteq_m G_j$$

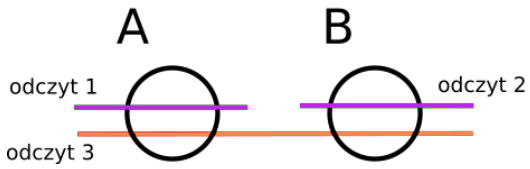
Powyższa definicja została dalej doprecyzowana o kryteria oceny potencjalnego rozwiązania, gdyż bez tego trudno mówić o optymalnym rozwiązaniu problemu. W wersji uproszczonej do wyszukiwania jednej sekwencji różnicowej, sekwencja ta musi spełniać warunki: bycia efektem asemblacji podzbioru odczytów, bycia różnicową, oraz bycia maksymalną pod względem długości.

Asemblacja przy użyciu grafu *de Bruijna* jest obecnie częstym rozwiązaniem, ponieważ użycie k-merów zamiast odczytów daje większą odporność na błędy sekwencjonowania i polimorfizm sekwencji a także jest korzystniejsze obliczeniowo. Jednak, jak pokazano w pracy, nie jest to dobre rozwiązanie dla asemblacji różnicowych kontigów, ponieważ wyznaczenie pokrycia ścieżki jest w tym grafie trudne obliczeniowo – połączenie ścieżek może dowolnie zwiększyć lub zmniejszyć pokrycie:

Obserwacja 2.1
Krotność zmiany $fc(A, B)$ *połączonych ścieżek* A *i* B *w grafie* de Bruijna *może dawać wartości większe lub mniejsze od* $fc(A)$ *i* $fc(B)$*.*

Ponadto zachodzi poniższa obserwacja, schematycznie zaprezentowana na rysunku 2.1.

Obserwacja 2.2
Dla dowolnego poziomu odcięcia $t > 1$ *i wielkości* k*-meru* k*, istnieje graf* de Bruijna *gdzie każdy wierzchołek ma* $fc < t$*, ale istnieje ścieżka o* $fc \geq t$*.*



Rysunek 2.1: Przykład grafu *de Bruijna*, w którym dla $FC = 2$ każdy z wierzchołków (A, B) ma $fc < FC$ a ścieżka $A \mapsto B$ ma $fc \geq FC$.

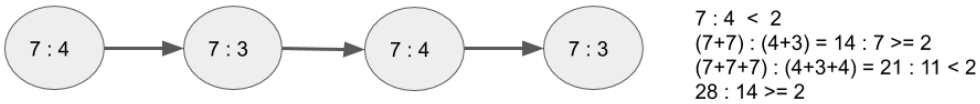
Stąd wniosek, że użycie tego typu grafu w sytuacji, w której konieczne jest śledzenie zliczeń odczytów przypisanych ścieżkom jest znacznie utrudnione. Inaczej sytuacja wygląda w grafie nałożeń odczytów. Liczba odczytów zmapowanych na ścieżkę powstałą z połączenia dwóch ścieżek to suma liczb odczytów zmapowanych do tych ścieżek. Ponadto zachodzi:

Obserwacja 2.3
Gdy połączymy ścieżki P *i* Q *w grafie nałożeń, to wartość funkcji* fc *będzie ograniczona przez jej wartości dla pojedynczych ścieżek:*

$$\min(fc(P), fc(Q)) \leq fc(P + Q) \leq \max(fc(P), fc(Q))$$

Z obserwacji 2.3 wynika, że w grafie nałożeń maksymalizacja krotności zmiany tworzonych kontigów będzie prowadziła do powstawania kontigów jednowierzchołkowych (które mogą być rozszerzone jeśli fc nie spadnie, ale i tak takie sformułowanie spowoduje otrzymanie krótszych kontigów).

Jednak nawet przy sformułowaniu problemu jako znalezienie najdłuższej ścieżki o zadowalającym fc, użycie grafu nałożeń sprawia kłopoty obliczeniowe, ponieważ mogą istnieć takie ścieżki, jak na rysunku 2.2, o fc naprzemiennie przekraczającym zadany próg. Nie wiadomo wtedy kiedy zakończyć poszukiwania kontigu, co powoduje konieczność szerokiego przeszukiwania grafu.



Rysunek 2.2: Przykład ścieżki o takich pokryciach (zaprezentowanych w wierzchołkach w formie *zliczenia w warunku pierwszym : w warunku drugim*), dla których trzeba sprawdzać wszystkie podścieżki żeby znaleźć optymalną.

Dlatego w dalszej części rozdziału 4 zaprezentowane są podejścia heurystyczne do budowy ścieżek, a wcześniej zanalizowane są parametry grafu nałożeń dla omawianych w tej pracy eksperymentalnych danych metatranskryptomicznych.

2.5 Wykorzystanie grafu nałożeń do asemblacji różnicowych kontigów

To, jak trudnym zadaniem w praktyce będzie asemblacja różnicowych kontigów w grafie nałożeń zależy od jego rozmiaru. Dlatego dla omawianych danych eksperymentalnych zostały w tej pracy zbadane różne parametry tego grafu, takie jak liczba wierzchołków i krawędzi oraz to, jak zmieniają się one wraz z upraszczaniem grafu, które jest ważną częścią asemblacji przy pomocy grafu nałożeń. Graf został zbudowany przy pomocy programu `SGA` (`String Graph Assembler`), a procedury upraszczania musiały zostać zmodyfikowane pod kątem asemblacji transkryptów. Upraszczanie naśladujące kroki wykonywane przez `SGA` zostało zaimplementowane w języku `Python`.

Nieoczyszczony graf okazał się bardzo duży, zawierał wiele wierzchołków super-repetytywnych (w uproszczeniu – mających ponad 128 sąsiadów), a z pozostałych po ich usunięciu krawędzi jedynie 9% miało orientację nieprawidłową dla odczytów z transkryptów, i zostało usunięte. Zmiany grafu w trakcie upraszczanie przedstawiono w tabeli 2.3.

	Wierzchołki:	Krawędzie:	Wyspy:	% odczytów:
Początkowo:	29 343 365	32 184 873	2 090 441	81.15%
- proste ścieżki		22 146 563		
(liczba ścieżek)	2 170 948			
- małe wyspy	2 601 781			
Po uproszczeniu:	4 595 021	10 038 310	96 100	75.02%
- końcówki	1 897 472			
- proste ścieżki (l. ścieżek)	315 075	518 514		
- małe wyspy	59 310			
Końcowo:	2 119 725	4 538 739	115 916	64.04%
SGA (10 iteracji)	1 423 561	7 749 584	59 528	
SGA (10 it., bez $\leftrightarrow$ i $\asymp$)	1 090 001	4 956 786	62,889	

Tabela 2.3: Rozmiary grafu nałożeń w kolejnych krokach heurystycznego upraszczania grafu: początkowo, czyli po wczytaniu (bez super-repetytywnych krawędzi), po uproszczeniu prostych ścieżek i po 1 cyklu usuwania końcówek, oraz rozmiary grafu po domyślnej procedurze `SGA` (10 iteracji i usuwanie bąbli), uruchomionej na zwykłem grafie i na grafie po usunięciu krawędzi prefiksowych i sufiksowych.

2.6 Heurystyki wyszukiwania różnicowych kontigów

W rozprawie zaproponowane są 3 proste „zachłanne" heurystyki wyszukiwania ścieżek, mające na celu znalezienie FC-różnicowych kontigów o możliwie największej długości:

1. Maksymalizująca długość (ozn. `longest`): dopóki są niewykorzystane wierzchołki:
   - wybieram najdłuższy (pod względem reprezentowanej sekwencji) wierzchołek jako początek ścieżki
   - dopóki to możliwe rozszerzam taką ścieżkę z dowolnego jej końca o najdłuższy wierzchołek

- wyjściowe ścieżki muszą zostać odfiltrowane by uzyskać tylko te zmienione

2. Maksymalizująca długość, ale różnicowa (ozn. `longestfc`): dopóki są niewykorzystane wierzchołki:
   - wybieram najdłuższy (pod względem reprezentowanej sekwencji) wierzchołek o $fc \geq FC$
   - dopóki to możliwe i wynikowa ścieżka ma $fc \geq FC$ rozszerzam taką ścieżkę z dowolnego jej końca o najdłuższy wierzchołek
3. Maksymalizująca krotność zmiany (ozn. `bestfc`): dopóki są niewykorzystane wierzchołki:
   - wybieram wierzchołek o największym $fc \geq FC$
   - dopóki to możliwe i wynikowa ścieżka ma $fc \geq FC$ rozszerzam taką ścieżkę z dowolnego jej końca o wierzchołek, który najmniej zmniejsza krotność zmiany.

Kontigi otrzymane różnymi heurystykami różniły się między sobą liczbą sekwencji, długością i pokryciem odczytów, chociaż różnice te nie były duże (patrz 2.4), dlatego do dalszych analiz zostały wybrane wyniki heurystyki `longestfc`.

Heurystyka	Liczba kontigów	sumaryczna długość (bp)	zliczenia	zliczenia (% odczytów)
longest ($fc \geq FC$)	182 973	85 289 444	42 317 834	23.3%
longest fc	213 507	99 725 607	49 295 380	27.2%
bestfc	220 888	100 319 295	46 086 786	25.4%
SGA	157 276	79 111 002	32 587 074	18.0%
SGA ($fc \geq FC$)	126 648	63 012 306	22 896 390	12.6%

Tabela 2.4: Porównanie kontigów otrzymanych przy pomocy różnych heurystyk

Porównanie kontigów poprzez mapowanie na nie odczytów (i na tej podstawie odfiltrowanie różnicowych) było konieczne aby porównać wyniki działania heurystyki z wynikami dla innych metod.

Pod względem długości oraz mapowalności metoda heurystyczna wypadła gorzej niż `MEGAHIT`, jednak mapowanie obu zbiorów kontigów przy pomocy `BLAST` [Altschul et al., 1990] na znane genomy pokazało, że część z tych sekwencji (większa w przypadku `MEGAHIT` niż heurystyki), szczególnie tych najdłuższych, może być wynikiem błędnej asemblacji (patrz 2.3).

3 Podsumowanie

W rozprawie szeroko omówione zostały zagadnienia badania wyników sekwencjonowania NGS danych metatranskryptomicznych (krótkich odczytów uzyskanych z RNA poddanego uprzednio rybodeplecji). Chociaż jest to rodzaj danych, który pozwala badać aktywność transkrypcyjną bakterii, to nie jest to jeszcze szeroko stosowany rodzaj eksperymentu. We współpracy z dr Iloną Grabowicz zaproponowana

`Kallisto` k=21

METODA	liczba kontigów	sumaryczna długość	zliczenia	zliczenia (% odczytów)
`longestfc`	156 385	73 638 169	59 820 097	33.02%
`SGA`	102 231	51 937 925	25 667 934	14.17%
`MEGAHIT`	148 475	139 477 791	102 960 602	56.83%

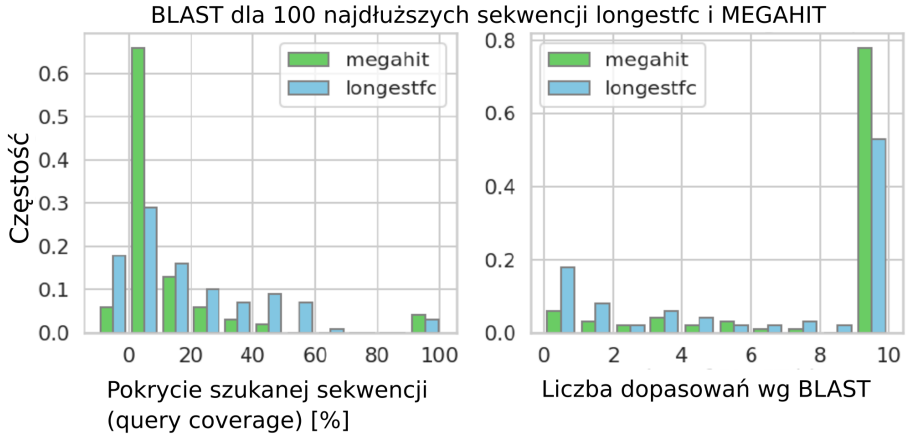
`bowtie`

METODA	liczba kontigów	sumaryczna długość	zliczenia	zliczenia (% odczytów)
`longestfc`	156 217	73 974 373	86 792 736	47.91%
`SGA`	104 549	52 976 914	49 683 447	27.42%
`MEGAHIT`	152 794	142 277 349	105 721 187	58.36%

`minimap`

METODA	liczba kontigów	sumaryczna długość	zliczenia	zliczenia (% odczytów)
`longestfc`	152 267	72 773 928	93 443 350	51.58%
`SGA`	101 778	51 892 087	68 950 876	38.06%
`MEGAHIT`	151 208	141 341 748	107 177 460	59.16%

Tabela 2.5: Liczba i pokrycie kontigów o $FC \geq 2$ na podstawie mapowania `kallisto`, `bowtie2` i `minimap2`.



Rysunek 2.3: Częstość znalezienia dopasowania o takim pokryciu przy pomocy programu `BLAST` dla 100 najdłuższych kontigów uzyskanych przez `MEGAHIT` oraz 2-różnicowych sekwencji uzyskanych przez heurystykę `longestfc`.

została wieloetapowa procedura używająca dostępnych narzędzi obejmująca asemblację, mapowanie, wybór genów i ich anotację funkcjonalną, oraz analizę różnicowej ekspresji. Procedura ta jest możliwa do wykorzystania w danych metatranskryptomicznych, ale zawiera wiele redundantnych kroków, a pozwoliła uzyskać jedynie niewielką liczbę statystycznie istotnie zmienionych genów (a dokładnie klastrów genów kodujących bardzo podobne białka).

Zaproponowane w dalszej części sformułowanie problemu asemblacji różnicowych kontigów pozwoliło opisać i zanalizować możliwość użycia do jego rozwiązania dwóch rodzajów grafów używanych do asemblacji – grafu *de Bruijna* i grafu nałożeń. Teo-

retyczne rozważania na temat struktury tych grafów i możliwości śledzenia w nich liczby odczytów pokazały, że graf *de Bruijna* jest raczej niemożliwy do wykorzystania w przypadku asemblacji różnicowych sekwencji. Graf nałożeń ma co prawda własność addytywności zliczeń, ale mogą w nim istnieć ścieżki różnicowe, których wykrycie może wymagać sprawdzenia wszystkich ścieżek. Ten fakt (a także zaprezentowane w pracy duże rozmiary grafu nałożeń dla danych eksperymentalnych) pokazuje, że potrzebne są podejścia heurystyczne.

Zaproponowany w pracy sposób wykorzystania `SGA` do budowy grafu nałożeń tak, aby możliwe było śledzenie liczby odczytów tworzących wierzchołki, a także pokazane heurystyki wyszukiwania różnicowych kontigów w tak przygotowanym grafie nałożeń, pozwoliły zaproponować zbiory sekwencji różnicowych, które są różne od sekwencji uzyskanych innymi metodami. Trudno w przypadku danych eksperymentalnych powiedzieć, która metoda jest lepsza. Pod względem liczby odczytów w różnicowych kontigach i pod względem sumarycznej długości `MEGAHIT` daje lepsze wyniki niż heurystyka `longestfc`, jednak porównanie mapowań `BLAST` sugeruje, że w sekwencjach zbudowanych przez `MEGAHIT` jest więcej sekwencji nieprawidłowych – chimerycznych sekwencji łączących prawdopodobnie poprawne fragmenty transkryptów. Ich zastosowanie powinno zatem zależeć od oczekiwań w stosunku do wyniku – zaprezentowana w tej pracy asemblacja z użyciem heurystyki lepiej sprawdzi się do uzyskania mniejszego zbioru sekwencji o większej pewności, a będzie gorsza w przypadku przygotowania szerszego zbioru, który podlega później dalszym analizom.

3.1 Dostępność

W przygotowaniu są publikacje dotyczące zagadnień opisanych w rozprawie:

- *Metatranscriptomic changes upon high-fat diet in a Down syndrome mouse model* autorstwa: Ilona E. Grabowicz, Julia Herman-Iżycka, Marta Fructuoso, Mara Dierssen, Bartek Wilczyński

- *application note* opisujący heurystyki asemblacji różnicowych kontigów w grafie nałożeń. Kod dotyczący tej części pracy jest dostępny pod adresem: `https://github.com/juliahi/diffcog`

Ponadto pod adresem `https://github.com/juliahi/kallisto` dostępna jest zmodyfikowana wersja `kallisto` pozwalająca uzyskać informację o mapowaniu bądź powodach niemapowania każdego analizowanego odczytu.

Literatura

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

[Bray et al., 2016] Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525.

[David et al., 2014] David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563.

[Foster and Neufeld, 2013] Foster, J. A. and Neufeld, K.-A. M. (2013). Gut–brain axis: how the microbiome influences anxiety and depression. *Trends in neurosciences*, 36(5):305–312.

[Goodrich et al., 2014] Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J. T., et al. (2014). Human genetics shape the gut microbiome. *Cell*, 159(4):789–799.

[Huerta-Cepas et al., 2017] Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., and Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggnog-mapper. *Molecular biology and evolution*, 34(8):2115–2122.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357.

[Li et al., 2015] Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676.

[Li, 2018] Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100.

[Lozupone et al., 2012] Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415):220.

[Myers, 1995] Myers, E. W. (1995). Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, 2(2):275–290.

[Peterson et al., 2009] Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C., et al. (2009). The nih human microbiome project. *Genome research*, 19(12):2317–2323.

[Pevzner et al., 2001] Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An eulerian path approach to dna fragment assembly. *Proceedings of the national academy of sciences*, 98(17):9748–9753.

[Pruitt et al., 2007] Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl_1):D61–D65.

[Schaeffer et al., 2017] Schaeffer, L., Pimentel, H., Bray, N., Melsted, P., and Pachter, L. (2017). Pseudoalignment for metagenomic read assignment. *Bioinformatics*, 33(14):2082–2088.

[Sgritta et al., 2019] Sgritta, M., Dooling, S. W., Buffington, S. A., Momin, E. N., Francis, M. B., Britton, R. A., and Costa-Mattioli, M. (2019). Mechanisms underlying microbial-mediated changes in social behavior in mouse models of autism spectrum disorder. *Neuron*, 101(2):246–259.

[Simpson and Durbin, 2012] Simpson, J. T. and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome research*, 22(3):549–556.

[Wood and Salzberg, 2014] Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultra-fast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):1–12.

[Zerbino and Birney, 2008] Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829.

[Zhu et al., 2010] Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic acids research*, 38(12):e132–e132.