UNIWERSYTET WROCŁAWSKI

WYDZIAŁ BIOTECHNOLOGII

ZAKŁAD MIKROBIOLOGII MOLEKULARNEJ

Przygotowanie bioinformatycznej bazy danych NGS dotyczących bakterii z rodzaju *Streptomyces*

Marek Skrzyński

Praca magisterska

Opiekun pracy : dr Agnieszka Strzałka

Wrocław, 2025 r.

STRONA NA PODZIĘKOWANIA

# Abstract/streszczenie

Wydaje mi się, że tak jak w przypadku pracy licencjackiej jak i sprawozdań itp. to napiszę na końcu.

Spis treści

[Abstract/streszczenie 3](#_Toc193205451)

[Wstęp/Rozwinięcie 6](#_Toc193205452)

[Streptomyces jako gatunek 7](#_Toc193205453)

[Gatunki modelowe *Streptomyces* 8](#_Toc193205454)

[Sekwencjonowanie genetyczne 9](#_Toc193205455)

[Bazy danych i narzędzia do analizy danych sekwencjonowania 11](#_Toc193205456)

[Cel pracy 12](#_Toc193205457)

[Materiały i metody 16](#_Toc193205458)

[Wyniki 18](#_Toc193205459)

[Dyskusja/Podsumowanie 19](#_Toc193205460)

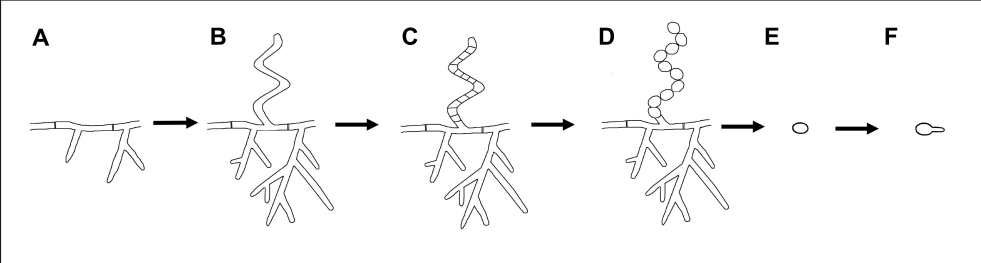
[Bibliografia 20](#_Toc193205461)

[Skróty i symbole 22](#_Toc193205462)

# Wstęp/Rozwinięcie

* Streptomyces coelicolor i venezuelae:
  + co to za bakterie
  + konkretniej co to za gatunki
  + ich zastosowanie w biotechnologii , z przykładami
  + ich znaczenie jako organizm modelowy, z przykładami
* NGS
  + Czym jest, do czego się wykorzystuje
  + Jakie znaczenie ma wykorzystywanie technik NGS w badaniu bakterii z gatunku streptomyces, z przykładami
  + jakie bazy danych już są – Strepdb, ArrayExpress, <https://imodulondb.org/index.html>
* Dane do aplikacji
  + Skąd zaciągnięte są dane/skąd je wzięliśmy
  + Problem z różnorodnością danych
    - Na czym polega
    - Że dane w suplementach
    - Różne rodzaje przeróbki od „raw” do publikowanych danych
  + Co tylko nam przyjdzie na myśl w kwestii „ktoś się może o to przyczepić”
  + Obróbka danych do programu, w kwestii, czemu były obrabiane
  + że są programy do analizy rna-seqów np. iDEP 2.01, https://github.com/federicomarini/GeneTonic
  + spróbuj wziąć surowe, wrzucić na idep a potem do swojego
* Aplikacja
  + Po co
    - Jaki jest jej cel
    - Czemu jest na nią zapotrzebowanie
  + Na co ma pozwalać użytkownikowi
  + Opis jak powstała, raczej po krótce
  + Nie wiem, tu coś powinno być chyba jeszcze, ale nie wiem co

## Streptomyces jako gatunek

Promieniowceto obszerny rząd organizmów prokariotycznych, a dokładniej są to bakterie gram dodatnie. Grupę tą cechuje wysoka zawartość par GC w materiale genetycznym, rozbudowany metabolizm drugorzędowy, jak i wysoki poziom zróżnicowania, często w obrębie jednego gatunki, pod kątem morfologicznym [1]. Do Promieniowców należą między innymi, bakterie rodzaju *Streptomyces*. *Streptomyces* to bardzo obszerny rząd bakterii liczący ponad 500 gatunków. Posiadają większy genom niż przeciętni reprezentanci prokariota, gdyż u *Streptomyces coelicolor* to aż 8,7 miliona par zasad, i tak jak innych reprezentantów Promieniowców, cechuje je wysoka zawartość par GC [2], [3]. Są to bakterie znajdowane głównie w glebie, a większość z przedstawicieli tego rzędu charakteryzuje złożony cykl życiowy składający się z fazy wzrostu wegetatywnego i sporulacji. Sposób rozmnażania się *Streptomyces* przyrównywany jest do tego u grzybów strzępkowych, lecz powstające u nich strzępko-podobne twory są mniej złożone morfologicznie i o mniejszym rozmiarze [4]. Umiejętność tworzenia strzępek wegetatywnych pozwala im na docieranie do składników odżywczych potrzebnych do ich przetrwania [5], [Rys.1]. Jednakże, gdy składników w okolicy bakterii zabraknie, wytwarzają one grzybnię powietrzną. W wyniku podziału strzępek powietrznych, powstają zarodniki dzięki czemu bakterie mogą szybko rozmieścić się na nowym terenie bogatym w składniki odżywcze i uniknąć niekorzystnych dla siebie warunków [6].

**Rysunek 1** Cykl życiowy Streptomyces. Schemat głównych etapów wzrostu: (A) Strzępki wegetatywne, (B) Wczesna sporulacja rozpoczynająca się od formowania grzybni powietrznej, (C) Powstanie niedojrzałych zarodników, (D) Dojrzewanie zarodników, (E) Uwalnianie dojrzałych zarodników, (F) Kiełkowanie w strzępki wegetatywne. Zaadaptowano z [7].

W wyniku tak rozbudowanych faz wzrostu powstaje wiele metabolitów drugorzędowych, które nie są niezbędne dla *Streptomyces* do wzrostu i przetrwania. Dzięki tej właściwości, ów rząd bakterii postrzegany jest jako źródło gamy różnorodnych związków naturalnego pochodzenia. Produkują one związki antybiotyczne takie jak streptomycyna, daptomycyna czy awermektyna [8]. Przy użyciu *Streptomyces* wyprodukowano również związki działające przeciw grzybom, stosowanym nawet przy infekcjach grzybiczych, takie jak amfoterycyna [8]. Przeprowadza się też badania nad klastrami genów, które potencjalnie mogłyby produkować metabolity drugorzędowe, lecz w warunkach laboratoryjnych są nieaktywne [9]. Również tematem ostatnio rozpatrywanym jest zastosowanie *Streptomyces* w heterologicznej produkcji białek, z nadzieją na lepszą wydajność niż u *Escherichia coli* [10].

## Gatunki modelowe *Streptomyces*

Główne dwa gatunki *Streptomyces*, szeroko stosowane w badaniach biotechnologicznych, to *Streptomyces* *coelicolor* (S.coelicolor) i *Streptomyces* *venezuelae* (S.venezuelae). S.coelicolor możemy nazwać klasycznym organizmem modelowym, ponieważ był to pierwszy gatunek *Streptomyces*, którego genom został w całości zsekwencjonowany, natomiast S.venezuelae jest nowym modelem ze względu na wyróżniającą się szybkość wzrostu i charakterystykę wzrostu w podłożu płynnym [2], [11]. S.coelicolor posiada około 7,8 tysiąca genów, a w całym genomie około 27 klastrów genów biosyntezy (BGC, z ang. biosynthesis gene clusters) kodujących białka, umożliwiające produkcję wyspecjalizowanych metabolitów. Bakteria ta szybko zyskała w oczach badaczy ze względu na kolor kolonii, które tworzy ten gatunek. Bakterie S.coelicolor wytwarzają związek zwany aktynorodyną, o kolorze niebieskim, który jest antybiotykiem kodowanym przez klaster genów *act*, a do tego jego barwa jest zależna od pH hodowli [11], [12]. S.coelicolor wytwarza również inne pigmenty w kolorach czerwonym, żółtym i szarym. Dzięki tej naturalnej właściwości, S.coelicolor, łatwo można odróżnić mutanty zablokowane na różnych etapach szlaku biosyntezy pigmentów [11].



**Rysunek 2** Zdjęcie kolonii S.coelicolor z grzybnią powietrzną i zabarwionymi na szaro zarodnikami. Ciemne zabarwienie wokół kolonii i krople na kolonii są wydzielona aktynorodyną. Zaadaptowano z [11].

S.coelicolor należy również do gatunków *Streptomyces* posiadających rozbudowany cykl życiowy, dlatego bada się regulację ekspresji genów i mechanizmów kontrolujących transkrypcję, w tym na przykład mutacje delecji genów związanych z rozwojem strzępek i fazy sporulacji. Jednym z najważniejszych mechanizmów regulacji transkrypcji u *Streptomyces* jest kontrola przez czynniki sigma. Czynniki sigma wchodzą w skład bakteryjnych polimeraz RNA i decydują one o specyficzności promotora, a w przypadku *Streptomyces* odgrywają również rolę w przekazywaniu sygnałów co ułatwia im przetrwanie w trudnych warunkach [13]. U S.coelicolor znajduje się około 64 czynników sigma, odgrywających różne role w cyklu życiowym tego organizmu. Można je podzielić na 4 grupy, różniące się ze względu na swoje funkcjonowanie i budowę [13]. Grupę pierwszą cechuje zdolność regulacji transkrypcji genów niezbędnych do funkcjonowania komórki, stąd tą grupę nazywa się konstytutywnymi czynnikami sigma. Grupa druga choć bardzo podobna w budowie do grupy pierwszej, odpowiada za regulację genów nieistotnych do wzrostu bakterii. Grupa trzecia, określana również jako alternatywne czynniki sigma, reguluje wiele szlaków metabolicznych w odpowiedzi na sygnały wewnątrz- jak i zewnątrzkomórkowe. Natomiast grupa czwarta odgrywa rolę w regulacji odpowiedzi na stres [13]. Czynniki sigma są niezwykle istotne w regulacji trankrypcji i funkcjonowania całej komórki, stąd przeprowadza się doświadczenia polegające między innymi na usuwaniu genów kodujących czynniki sigma. Usunięcie czynnika sigma B, należącego do grupy trzeciej wykazało, że jest on głównym regulatorem, który aktywuje inne czynniki sigma w kaskadzie, a tym samym kontroluje proces różnicowania się komórek oraz odpowiedź oksydacyjną S.coelicolor [14].

*Streptomyces* posiadają również setki czynników traksrypcyjnych, innych niż czynniki sigma, które modulują ekspresję genów na poziomie globalnym komórki, jak i takich, które odpowiadają za specyficzne szlaki biosyntezy. Czynniki transkrypcyjne należące do globalnych zwykle regulują ekpresję nawet całego klastra genów, na którym mogą się znajdować [15]. Do takich należy między innymi gen Bld. Mutanty *bld*,bez genów regulatorowychBld, które odpowiedzialne są za powstawanie strzępek powietrznych, nie posiadają na koloniach „włosków” typowych dla kolonii S.coelicolor, stąd z ang. „bald” otrzymuje się fenotyp „łysy”, łatwy do rozpoznania. Innym genem działającym szeroko na regulacje procesów komórkowych jest Whi, odpowiadających za podział strzępek powietrznych w spory. W mutantach *whi* S.coelicolor, nie zauważymy szarego pigmentu charakterystycznego dla dojrzałych spor. Z kolei do genów o małym zasięgu regulacji należą, między innymi ActIIORF4, czy RedD, które odpowiadają jedynie za regulacje biosyntezy konkretnych pigmentów występujących w S.coelicolor. Jednak poszerzanie wiedzy powoduje, że tak proste zależności są obalane. Czasem dochodzi do regulacji krzyżującej się między różnymi genami i spotkać się możemy z sytuacją, w której gen regulujący specyficzny szlak, reguluje również gen o szerszym zakresie działania [15].

Innym czynnikiem regulującym ekspresję genów są białka związane z organizacją nukleoidu (Nucleoid-associated proteins, NAP). Odgrywają one kluczową rolę w rearanżacji DNA, wiązaniu go i stabilizowaniu domen powstałych z materiału genetycznego [16]. Te funkcje białek NAP powodują, że wpływają one na dostępność promotorów dla czynników transkrypcyjnych i polimerazy RNA. Do białek NAP należą między innymi, białka z grupy homologów HU, HupA i HupS, białko regulujące segregację chromosomu ParAB, czy białko z grupy kondensyn, SMC [16].

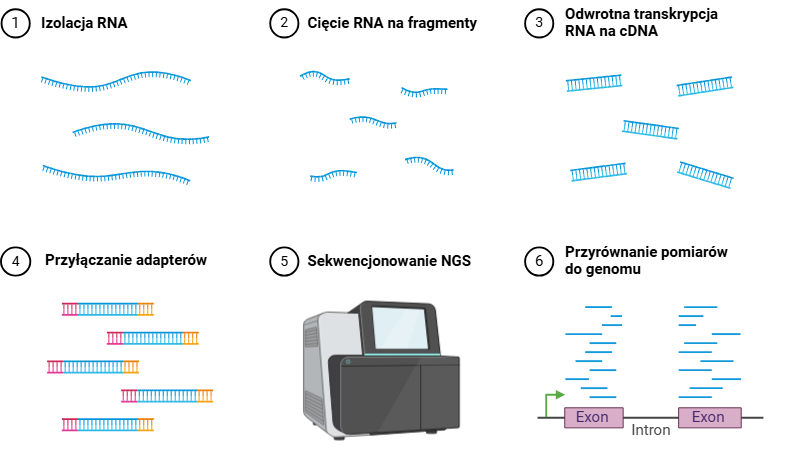
Dzięki wszystkim powyższym właściwościom S.coelicolor były używane jako organizm modelowy dla *Streptomyces* przez lata, niestety ma on swoje ograniczenia, między innymi, jak wiele innych gatunków tego rzędu, do jego sporulacji dochodzi tylko na podłożu stałym do tego w sposób asynchroniczny, a w mediach płynnych tworzy duże skupiska strzępek mycelialnych [11]. Wynikiem tego jest coraz większe zainteresowanie w S.venezuelae jako reprezentancie *Streptomyces*. Jest to gatunek o trochę mniejszym genomie niż S.coelicolor, posiadającym 7,1 tysiąca genów kodujących białka, a w całym genomie około 34 BGC. Około 85% genów kodujących białka to ortologi genów S.coelicolor, ale w przeciwieństwie do S. coelicolor, S.venezuelae szybciej i łatwiej osiąga synchronizacje rozwoju w mediach płynnych oraz przechodzi w nich cały cykl życiowy. Oprócz podstawowych faz wzrostu takim jak wegetatywny i sporulacyjny S.venezuelae wykazuje też tryb wzrostu eksploracyjnego, charakteryzującego się dynamicznym rozprzestrzenianiem się strzępek bakterii po podłożu, na przykład, w wyniku niedoboru składników odżywczych. W ten sposób bakteria może szybko rozprzestrzenić się w pobliskim środowisku bez konieczności sporulacji [17]. Dzięki tym cechom S.venezuelae jest przewodnim gatunkiem w badaniu cyklu życiowego bakterii *Streptomyces*, ponieważ z hodowli płynnych z łatwością można wykonać preparaty do badań mikroskopowych w tym do badań z użyciem mikroskopii fluorescencyjnej [18]. Charakterystycznym pigmentem dla S.venezuelae jest zielony barwnik powstający w dojrzałych zarodnikach [18]. Produkują one również inne związki, które są nieobecne u S.coelicolor, między innymi chloramfenikol, który jest antybiotykiem o bardzo szerokim działaniu, działającym poprzez inhibicje syntezy białek poprzez wiązanie z podjednostką 50S rybosomu bakteryjnego [19]. Do tego większość metod stosowanych do modyfikacji genetycznych S.coelicolor, działa również na S.venezuelae

**Rysunek 3** Kolonia Streptomyces venezuelae podczas sporulacji. Charakterystyczne zielone pokrycie spowodowane obecnością pigmentu o tym kolorze w dojrzałych zarodnikach [18].

Oba te gatunki cechuje rozbudowany cykl życiowy, wiele szlaków biosyntezy metabolitów drugorzędowych i klastrów biosyntezy genów. Zatem z tych cech wynika potencjał do usprawnienia wydajności produkcji antybiotyków, jak i ich samych, poprzez manipulacje BGC, które produkują naturalnie antybiotyki, oraz inżynierię szlaków metabolicznych i syntezę białek rekombinowanych pozwalających na nowe terapie. Od kilku lat kładzie się nacisk na zgłębianie działania szlaków biosyntezy w *Streptomyces*, a dokładniej, jak konkretnie wpływają one na siebie nawzajem i jakie modyfikacje możemy wprowadzać, aby usprawnić funkcjonowanie, bądź produkcje niektórych metabolitów. Oprócz tego bada się jak powiązane są te szlaki z procesami towarzyszącymi fazom wzrostu w warunkach hodowlanych, często pod wpływem różnego rodzaju czynników stresowych. Do takich rodzajów badań można stosować metody z rodzaju technik sekwencjonowania nowej generacji takie jak RNA-seq, ChIP-seq i mikromacierze.

## Sekwencjonowanie genetyczne

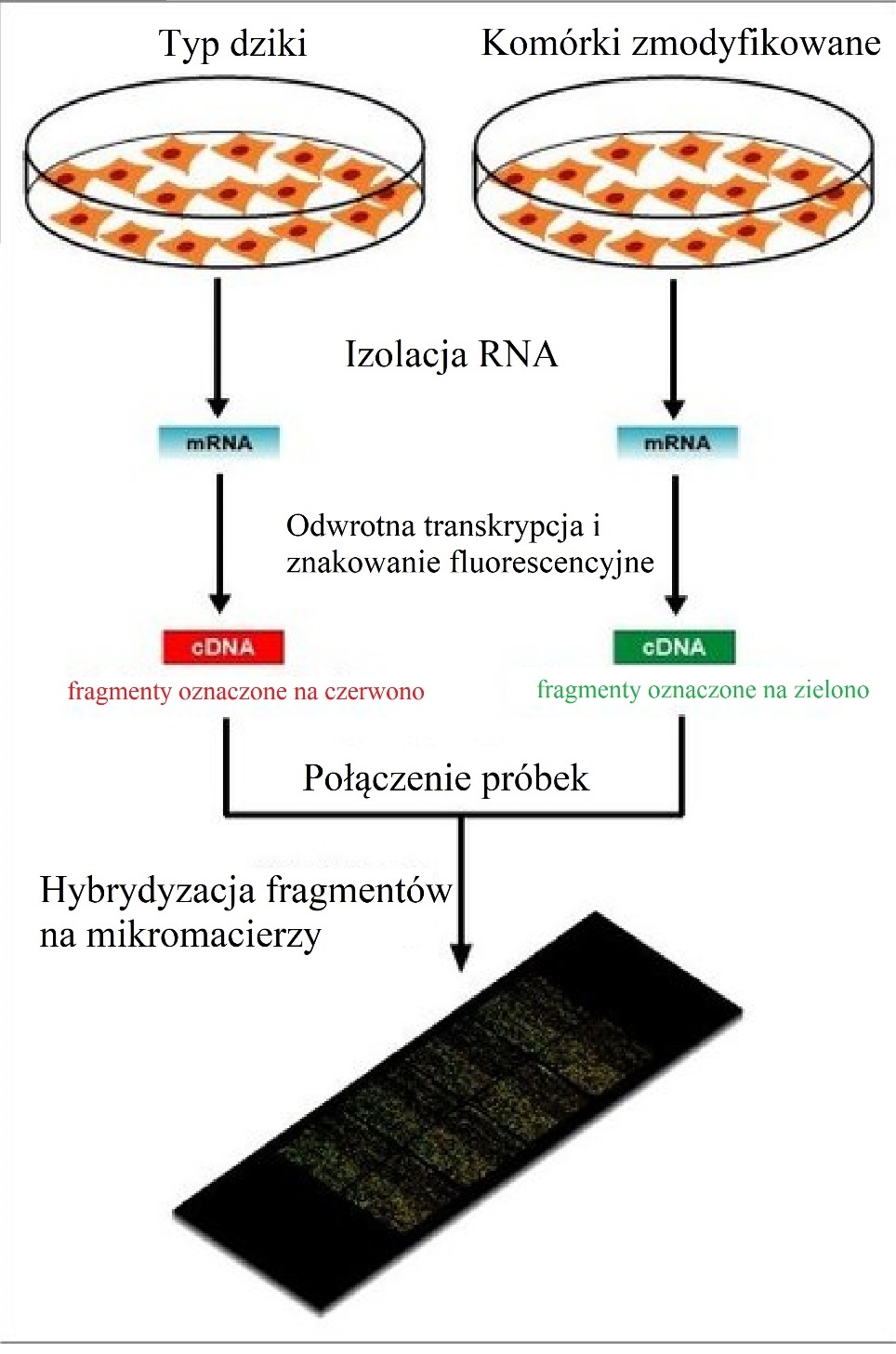
Od zsekwencjonowania genomu ludzkiego techniki odczytywania sekwencji materiału genetycznego rozwijały się w gwałtownym tempie. Do zapisania jak wygląda genom ludzi zastosowane zostało sekwencjonowanie pierwszej generacji, a dokładnie sekwencjonowanie metodą Sanger-a. Największym problemem tej techniki w na początku XXI wieku była jej przepustowość. Stąd pojawiła się potrzeba powstania nowych technik, które oprócz umożliwiania wyższej przepustowości badań, byłaby też bardziej opłacalna od strony finansowej. Dlatego już wtedy zaczęły pojawiać się techniki drugiej generacji sekwencjonowania, zwane teraz sekwencjowaniem nowej generacji (New generation sequencing, NGS). Sekwencjonowanie pierwszej i drugiej generacji, działają na podobnych zasadach, lecz NGS zdecydowanie kładzie nacisk na dużo większą ilość możliwych odczytów w jednym momencie. Do tego metoda Sangera opierała się na odczytaniu sekwencji jednej nici, a metody NGS pozwalają na przeanalizowanie całego genomu na raz, w podobnej ilości czasu. Przede wszystkim w NGS używa się nukleotydów oznakowanych znacznikami fluorescencyjnymi, o 4 różnych kolorach odpowiadających zasadom azotowym A, C, T i G. Wszystkie nukleotydy używane w NGS są oznakowane w odróżnieniu od metody Sangera [20]. Do przeprowadzenia NGS przygotowuje się preparat DNA, oczyszczony z RNA i rRNA, a następnie DNA jest cięte na krótkie fragmenty. Następnie do fragmentów przyłączane są adaptery. Fragmenty DNA są rozdzielane na pojedyncze nici. Następnie próbkę przenosi się na kasetę przepływową, na której szklanej powierzchni znajdują się krótkie nukleotydy komplementarne do adapterów na fragmentach DNA. Pojedyncza nić DNA przyłącza się do oligonukleotydu na kasecie. Do fragmentów dołączane są nukleotydy fluorescencyjne. Między każdym dołączeniem nukleotydu dokonuje się pomiaru fluorescencji i zahamowania emisji przyłączonego nukleotydu. W ten sposób powstaje szereg kolejnych odczytów, z których jesteśmy w stanie odczytać jaka jest sekwencja danego fragmentu. W ten sposób jesteśmy w stanie odczytać miliony par zasad podczas jednego przebiegu NGS.

Sekwencjonowanie RNA (RNA-seq), umożliwia analizę transkryptomu, czyli dokładniej poziomu ekspresji wszystkich genów w różnych warunkach, takich jak, różne fazy wzrostu czy obecność czynników stresowych. Istnieją różne systemy dostępne na rynku, takie jak FLX, Illumina czy AB SOLiD system, umożliwiające wykonywanie pomiarów RNA-seq, lecz operują one na podobnych założeniach [21]. Systemy te różnią się głównie tym jakiej długości fragmenty DNA są przyjmowane do sekwencjonowania. Głównym założeniem RNA-seq jest pomiar ekspresji, więc nie operuje się tu na próbkach DNA. Z kilku powtórzeń biologicznych bądź technicznych pobierany jest materiał genetyczny RNA, który powstaje w wyniku ekspresji DNA obecnego w bakterii. RNA jest dzielone na fragmenty, które są transkrybowane na dwuniciowe cDNA i dopiero wtedy przeprowadza się ich sekwencjonowanie. Pozyskane krótkie fragmenty są sekwencjonowane równolegle, w efekcie czego otrzymujemy duże ilości krótkich odczytów.

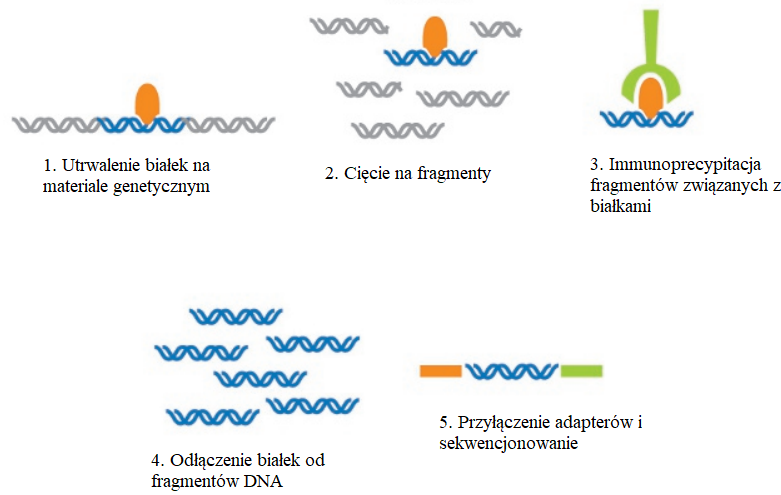
**Rysunek 4** Schemat zasady działania RNA-seq. Zaadaptowano z [22]

Stosując metodę RNA-seq możemy dokonać pomiaru nie znając genomu wyjściowego, jednakże genom S.coelicolor i S.venezuelae jest znany, więc głównie chodzi o zaobserwowanie liczby odczytów danych genów, ponieważ dzięki takiemu podejściu możemy otrzymać różnice w ekspresji danych genów. Zwykle ta różnica jest określana jako fold change (FC), czyli stosunek ekspresji genu w dwóch warunkach na przykład między typem dzikim bakterii i mutantem z delecją konkretnego genu. Parametr FC często podawany jest w postaci logarytmu o podstawie dwa i zapisywany jako „logFC”. Ważnym parametrem też jest wskaźnik fałszywych odczytów (FDR z ang. false discovery rate). FDR liczony jest na podstawie wartości „p” dla pomiarów i dostarcza informacji o tym jaka część wyników istotnych statystycznie może być fałszywie pozytywna, dlatego często analizuje się wyniki w oparciu o wartość FDR mniejszą od 0,05. RNA-seq można zastosować do porównania jak delecja danego genu wpłynie na geny z nim sąsiadujące, bądź na inne powiązane z nim geny w porównaniu do typu dzikiego tej samej bakterii [23], [24]. Można też przeprowadzić pomiary w różnych fazach wzrostu bakterii i porównać, jak wygląda ekspresja genów w różnych etapach cyklu życiowego [25].

Mikromacierze wymagają od nas znajomości genomu badanej bakterii. Technika ta jest podobna do RNA-seq, lecz jej zasada działania jest inna. Do przeprowadzenia takiego badania potrzebna nam jest macierz, na której w dołkach znajdują się fragmenty materiału genetycznego odpowiadające konkretnym genom bądź regionom kodującym. Z dwóch prób pobiera się RNA i przetłumacza na cDNA poprzez zastosowanie odwrotnej transkrypcji. Następnie cDNA poddaje się fragmentacji i oznacza fluorescencyjnie różnymi znacznikami. Nakładane na macierz próbki, z obu prób biologicznych na raz, będą hybrydyzować z fragmentami w dołkach macierzy, jeśli będą do nich komplementarne, a reszta próbki jest odmywana. W ten sposób z każdego dołka uda nam się odczytać sygnał fluorescencyjny, a na podstawie obliczeń jest się w stanie ocenić jaka jest ekspresja różnicowa genów między badanymi bakteriami. Jest to metoda szybsza niż RNA-seq, ale mniej czuła, lecz stosuje się ją do podobnych eksperymentów jak RNA-seq. Również jak w RNA-seq, wylicza się tu parametry FC, jak i FDR. Technikę można stosować do porównania mutantów *Streptomyces* ze zmienionymi sekwencjami genów regulatorowych z typem dzikim, w celu zbadania znaczenia danych genów w procesach zachodzących podczas sporulacji [26], bądź do bardziej globalnego porównania zmian ekspresji w genomie w momencie hodowania bakterii w różnego rodzaju pożywkach, na przykład stałej i płynnej [27].



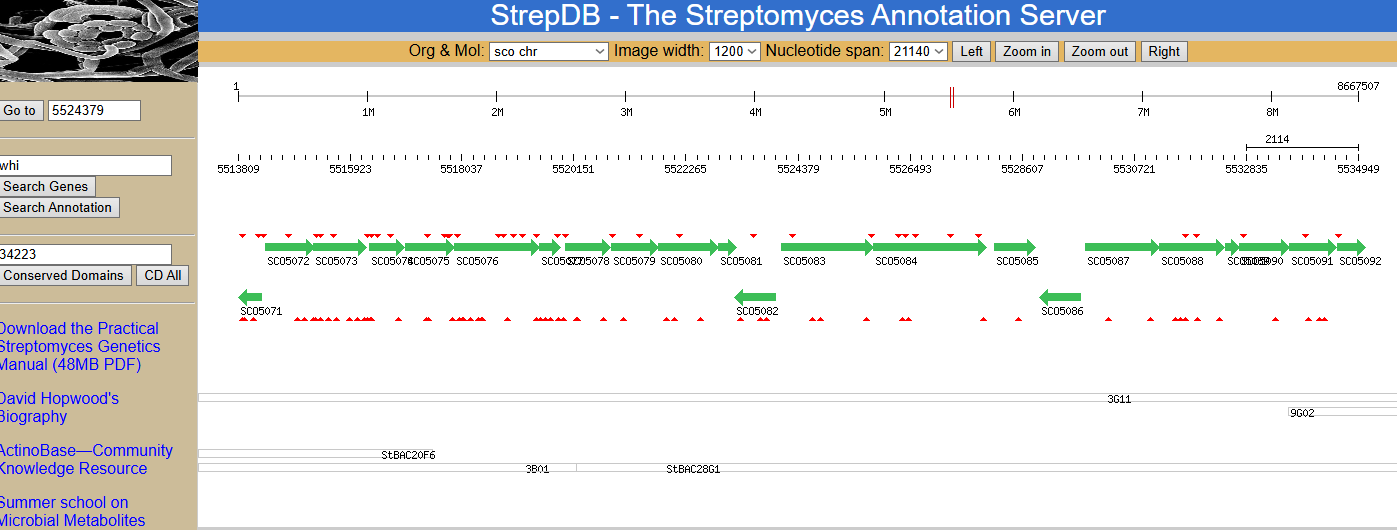
**Rysunek 5** Schemat procedury zastosowania mikromacierzy.

ChIP-seq stosowany jest do identyfikacji miejsc wiązania białek regulatorowych takich jak czynniki transkrypcyjne, do chromosomu. Komórki traktowane formaldehydem w celu zatrzymania białek przy DNA, a następnie materiał tnie się na krótkie fragmenty. Przeprowadzana jest immunoprecypitacja z przeciwciałami specyficznymi dla danego białka, wtedy oznaczone fragmenty są odzyskiwane i oczyszczane z przyłączonych białek, po czym poddawane sekwencjonowaniu. W ten sposób otrzymujemy zsekwencjonowane tylko fragmenty, do których przyłączone było interesujące nas białko. Dzięki zastosowaniu techniki ChIP-seq możemy określić, jak zachodzą procesy związane z odczytywaniem materiału genetycznego, ale też jak wygląda segregowanie materiału w momencie podziału komórek, czy ich rozwoju. Można określić, gdzie przyłączają się białka NAP, takie jak HupS, gdy zachodzi rearanżacja materiału *Streptomyces* w momencie powstawania spor, a dokładnie podczas podziału komórek na spory [28]. Używa się też do badania regulonów czynników transkrypcyjnych w bakteriach, jak na przykład zidentyfikowanie genów kontrolowanych przez konkretne białko [29].

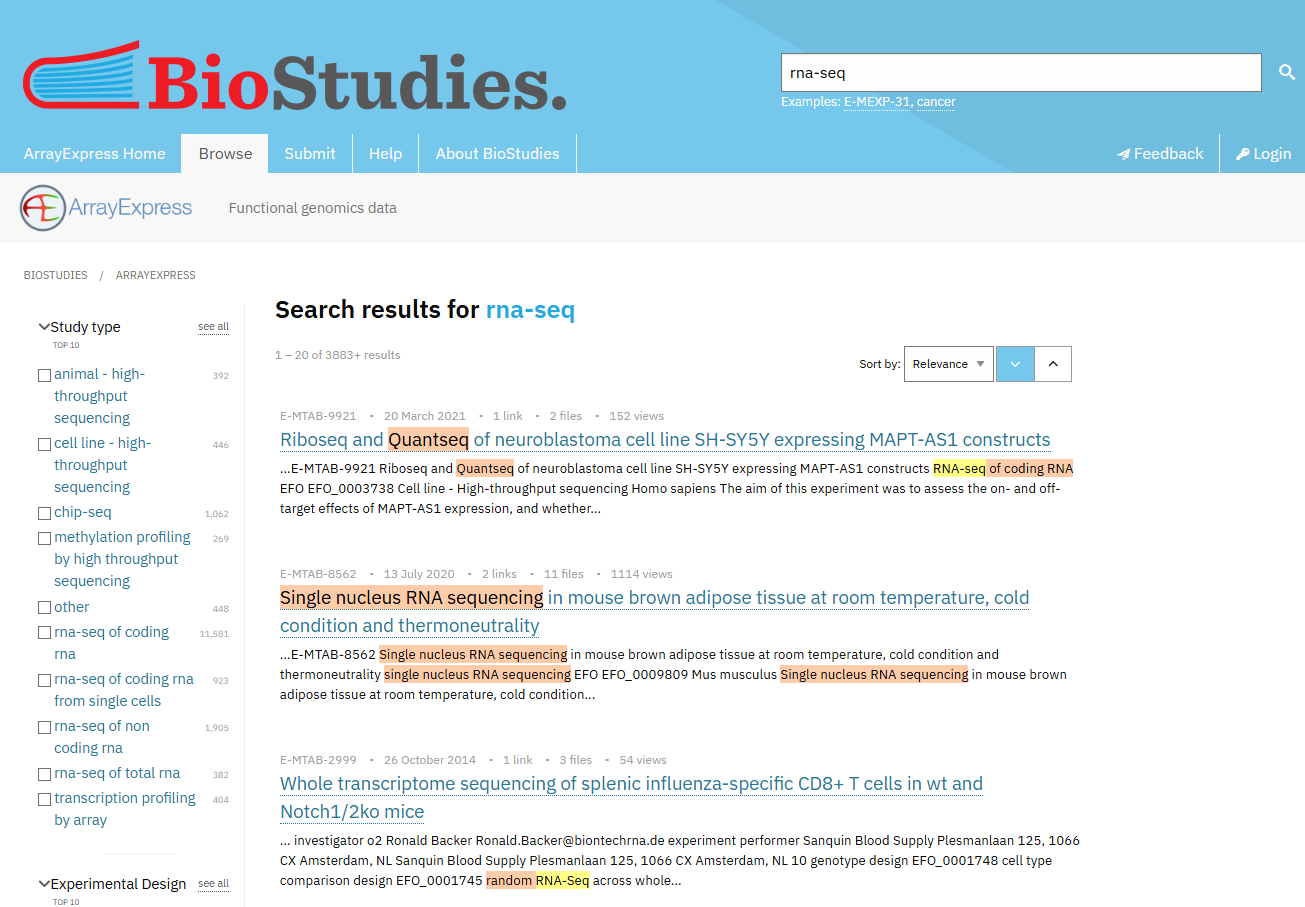
**Rysunek 6** Schemat przygotowania preparatu w ChIP-seq. Zaadaptowano z [30]

## Bazy danych i narzędzia do analizy danych sekwencjonowania

Ponieważ istnieje tyle danych dotyczących powyżej wspomnianych zagadnień, powstały strony umożliwiające wgląd w istniejące dane pochodzące z sekwencjonowania bakterii *Streptomyces*, zawierające cały genom konkretnych gatunków, jak i wyniki z różnego rodzaju eksperymentów. Do najczęściej używanych należą bazy danych takie jak StrepDB, ArrayExpress i iModulons, oraz narzędzia bioinformatyczne do analizy danych takie jak iDEP. Oferują one szeroką gamę możliwości wglądu w genom *Streptomyces*, lecz każda z nich ma też swoje ograniczenia.

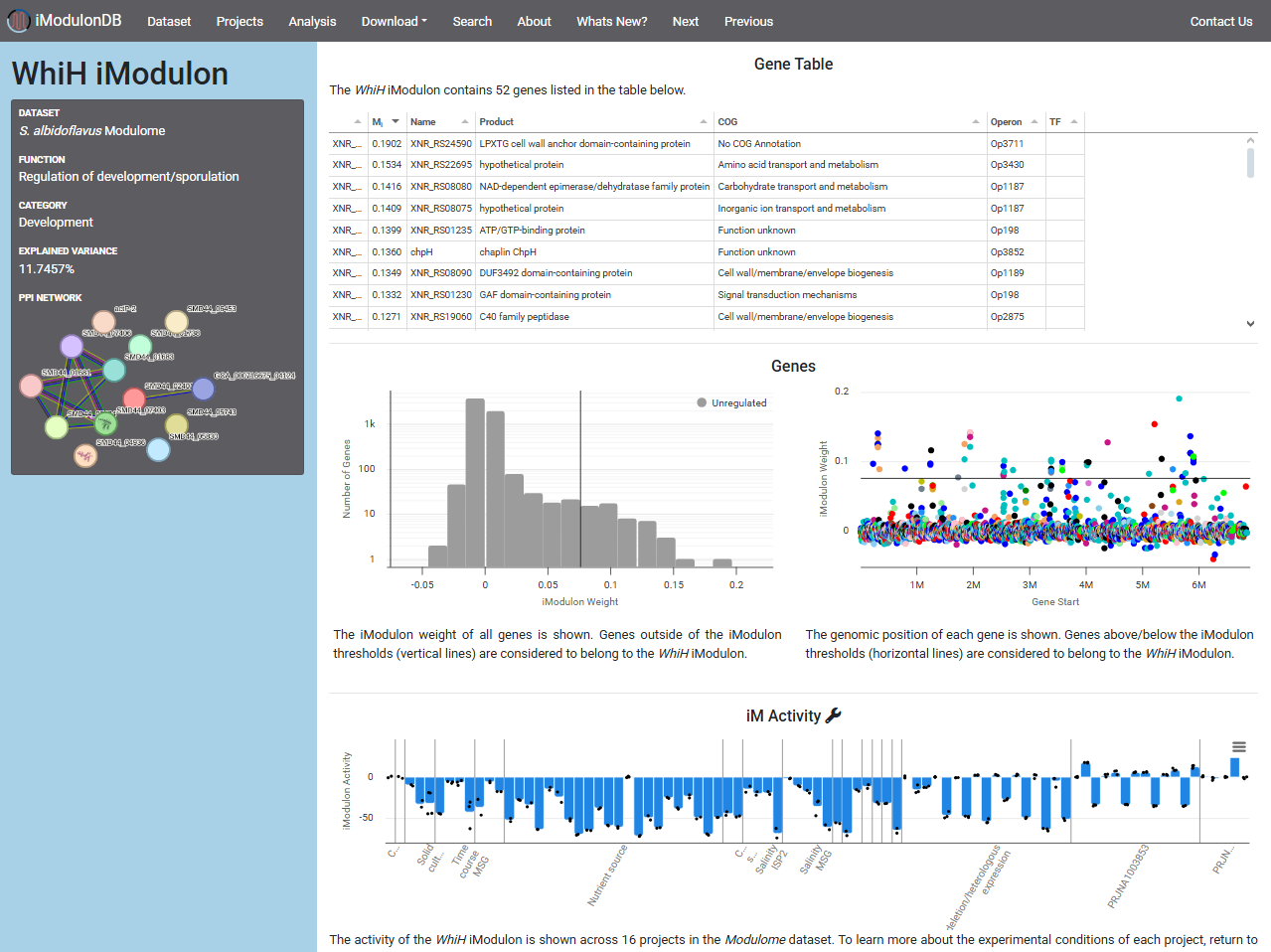
StrepDB oferuje szeroki zakres zasobów związany z badaniami nad *Streptomyces*, jest to baza zawierająca genomy różnych gatunków, metabolitów, BGC, genów czy szlaków biosyntezy. Strona udostępnia narzędzia do analizy danych takie jak BLAST jak i dokładne opisy genów znajdujących się na wykresach i domen. Natomiast na StrepDB nie dostaniemy możliwości obejrzenia najnowszych danych wraz poziomami ekspresji, a tym bardziej nie znajdziemy na niej opcji obejrzenia własnych danych, raczej służy ona do dokładnego wglądu w istniejące już, pełne, genomy gatunków *Streptomyces*.

**Rysunek 7** Klaster biosyntezy aktynorodyny wyświetlony w bazie StrepDB

ArrayExpress jest bazą danych, która udostępnia dostęp do danych z eksperymentów z zakresu genomiki, transkryptomiki i innych badań omicznych, w tym danych z RNA-seq, czy mikromacierzy [31]. Odnajdziemy tam szeroki zakres danych eksperymentalnych z różnych dziedzin i pozyskanych z różnych organizmów. Pozwala ona również na udostępnianie własnych danych do wglądu przez innych użytkowników, przeglądanie danych już

**Rysunek 8** Wyniki wyszukiwania dla hasła „rna-seq” w bazie ArrayExpress

znajdujących się w bazie danych jak i wyszukiwanie danych według różnych kryteriów, takich jak typ danych, organizm, technika, rodzaj próbki, oraz warunki eksperymentalne. Istnieją również narzędzia wyszukiwania umożliwiające łatwe odnalezienie konkretnych danych. Dane również zawierają opisy z jakich eksperymentów pochodzą. Strona ta nie oferuje jednak wglądu do danych w sposób graficzny, są one tam dostępne do pobrania i zgłębienia wyników samemu, co może okazać się skomplikowane dla niektórych użytkowników ze względu na różnorodność formatów.

iModulonDB to baza danych dedykowana dla *Streptomyces* jak i innych organizmów bakteryjnych, oferująca informacje o regulacji genów i szlaków metabolicznych [32]. Jest ona skoncentrowana na tym, jak regulacja ekspresji genów wpływa na procesy biologiczne, takie jak wzrost, produkcja metabolitów czy reakcje na stres i inne bodźce. Oferuje ona estetyczny graficzny wgląd w regulony i szlaki w niektórtch bakteriach z dokładnymi opisami danych i możliwością modyfikacji wyświetlanych wykresów. Niestety nie znajdziemy na niej informacji dotyczących *Streptomyces* *coelicolor* jak i *venezuelae*, w naszym przypadku, gatunków nas interesujących.

**Rysunek 9** Prezentacja wyników dla WhiH *Streptomyces albidoflavus* w bazie iModulonDB

iDEP jest stroną, która funkcjonuje raczej jako narzędzie bioinformatyczne aniżeli baza danych [33]. Na iDEP jesteśmy wstanie przeanalizować dane z RNA-seq i mikromacierzy, a następnie wyświetlić przeanalizowane dane w formie wykresów i tabel, które są modyfikowalne. Jest to narzędzie dostępne online bez potrzeby pobierania oprogramowania i umożliwia użytkownikowi analizę różnicowej ekspresji genów oraz powiązanych szlaków metabolicznych i funkcjonalnych. Niestety, jak już zostało wspomniane, iDEP nie jest bazą danych, co oznacza, że poza plikiem „demo” nie znajdziemy tam innych danych, które będziemy mogli obejrzeć, jedynie te, które sami załadujemy na stronę.

# Cel pracy

DO POPRAWY Stąd zaistniał pomysł na stworzenie własnej bazy danych, która będzie spełniała wszystkie interesujące nas warunki. Głównym jej celem byłoby wyświetlanie danych pochodzących z RNA-seq, mikromacierzy i ChIP-seq z różnych eksperymentów na S.coelicolor i S.venezuelae, ze względu na rosnące zainteresowanie takimi danymi. Chcieliśmy, żeby program który powstanie, nie tylko gromadził interesujące użytkowników dane, ale udostępniał je również do wglądu z możliwością analizy różnicy w ekspresji między różnymi eksperymentami. Oprócz tego, ważne było, aby użytkownik mógł załadować własne dane, aby móc je porównać z istniejącymi. Takie narzędzie, łączyłoby cechy typowej bazy danych i narzędzi bioinformatycznych do obrazowania wyników. Chcieliśmy by nawet po zakończeniu głównego projektu związanego z pracą magisterską, można było do programu dodawać nowe dane i poszerzać dostępną w nim bazę, aby pozostawał użytkowy. Przede wszystkim wyniki powinny być zapisane tak aby użytkownik wiedział z jakimi eksperymentami ma do czynienie i pełną swobodę w tym co ma być wyświetlone. Nie chodzi tu tylko o to jakie dane, ale też jakie wykresy powinny być widoczne. Oprócz tego powinien widoczny być genom jako referencja do wyświetlanych danych, ponieważ nie wszystkie geny są odnotowane we wszystkich eksperymentach. Dostępna powinna być opcja poruszania się po wyświetlanych danych, ponieważ niemożliwe będzie wyświetlenie całego genomu *Streptomyces*. Do tego możliwość filtracji wyników na podstawie konkretnych kryteriów. Dodatkową opcją może być opcja wyświetlania porównania między konkretnymi danymi, w celu odnotowania podobieństw między zbiorami. Jako, że to program wyświetlający dane głównie w sposób graficzny, należy dodać opcję pobrania wyświetlonych wyników, w razie chęci zamieszczenia ich we własnej publikacji. Ze strony twórcy udostępniona również będzie instrukcja obsługi aplikacji, a także opis wszystkich danych zebranych w jej bazie.

Dane zostały zgromadzone z publikacji, które wybierane były na podstawie opisywanych w nich eksperymentów i interesujących nas gatunków. Wszystkie dane załączone w aplikacji pochodzą z eksperymentów dotyczących *Streptomyces* *coelicolor* i *Streptomyces* *venezuelae*, i wykonywanych na nich badań z zakresu NGS. Oczywiście dane te są dostępne dla każdego, lecz problem techniczny pojawia się w momencie, gdy próbujemy porównywać ze sobą wyniki zaprezentowane w artykułach. Często autorzy skupiają się tylko na konkretnych genach interesujących ich w swoich badaniach, dlatego w samym tekście publikacji trudno o szczegółowy wynik z całego sekwencjonowania. Bardziej rozbudowane dane można znaleźć w suplemencie znajdowanych publikacji. Powoduje to, że dla badacza, skomplikowane może się okazać zestawienie danych występujących w tak dużej ilości różnych formatów. Spotkać można się z wynikami w postaci tabeli w arkuszu kalkulacyjnym Excel, tabeli zapisanej w formacie PDF, z której ciężko nawet o przekopiowanie danych do innego oprogramowania z zachowaniem ich sensu, czy też tabele w programie Word. Zdarzają się również dane w plikach o rozszerzeniach, których nie jest się w stanie otworzyć bez zainstalowania dodatkowych programów, ewentualnie część jest otwieralna za pomocą edytora tekstu obecnego na systemach komputerowych, ale tu też może się okazać, że plik jest zapisany w sposób binarny i otwarcie go jako tekst nie umożliwi nam do niego wglądu. Niestety z takim kompletowaniem danych pojawia się problem, gdy przychodzi do ich porównywania. Autorzy praktycznie nie udostępniają danych, przed ich przeliczeniem przez programy do tego służące, czyli w tak zwanej wersji surowej („raw”), dlatego nie jesteśmy w stanie zagwarantować, że wszystkie zebrane do programu dane zostały znormalizowane w ten sam sposób, a nawet możemy stwierdzić, że na pewno nie były. Jest to problem, na który nie ma rozwiązania, jeśli chcemy zbierać dane dostępne publicznie.

TO DO MATERIAŁY I METODY Program powstał przy użyciu języka programowania R [1, a dokładniej oprogramowania RStudio [1. R to język programowania i środowisko obliczeniowe głównie do analizy statystycznej, wizualizacji danych oraz obliczeń naukowych. Jest on szeroko stosowany w bioinformatyce. Pozwala on na obsługę dużych zbiorów danych, dzięki czemu idealnie nadaje się do pracy z danymi eksperymentalnymi, jak w tym przypadku, danych z sekwencjonowania. Pozwala on na tworzenie zaawansowanych wykresów i wizualizacji danych, szczególnie dzięki pakietom, czyli dodatkowym zestawom funkcji, do niego dostępnych, takich jak ggplot2 [1. Język R jest też oprogramowaniem open-source i ma ogromną społeczność rozwijającą dodatkowe pakiety, a ze względu na zainteresowanie ze strony dziedziny bioinformatyki, powstają takie, które są dedykowane do danych biologicznych, na przykład gggenes [1, czyli rozszerzenie ggplot2, umożliwiające tworzenie wykresów strzałkowych obrazujących geny. Natomiast RStudio to zintegrowane środowisko programistyczne dla R, które ułatwia pracę w tym języku. Pozwala ono na lepszą organizację kodu i danych, autouzupełnianie kodu, debugowanie i podgląd danych czy zarządzanie pakietami. Nasza aplikacja powstała przy użyciu pakietu Shiny [1. Jest to pakiet, który powstał, by pozwolić na tworzenie interaktywnych aplikacji webowych, czyli takich, które uruchamia się w przeglądarce. Shiny ułatwia tworzenie aplikacji bez potrzeby znajomości HTML czy JavaScript. Aplikacje powstałe przy użyciu tego pakietu pozwalają na wgląd do udostępnionych danych i interakcje z nimi, w tym ich wizualizację.

Tak jak człowiek nie jest w stanie łatwo poradzić sobie z wieloma plikami w różnych formatach, podobnie jest z programem komputerowym. Aby aplikacja w Shiny w R działała, dane muszą być ujednolicone w jednym formacie, najlepiej zgodnymi z zasadami „tidy data” [1. Dane „tidy” charakteryzują się głównie trzema zasadami, każda kolumna to jedna zmienna (temperatura, czas, kontrast), każdy wiersz to jedna obserwacja (gen, chromosom), każda komórka zawiera jedną wartość. Mimo tego co by się mogło wydawać, jak może wyglądać format „tidy”, okazuje się, że zdecydowana większość danych, które można znaleźć w suplementach publikacjach czy bazach danych nie są w tym formacie. Dlatego należało przeformatować wszystkie zebrane dane na „tidy”, a aby to osiągnąć należało wymyśleć ujednolicony układ/format danych do aplikacji.

# Materiały i metody

* Wszystko tu będzie obrzucone niezliczoną ilością screenów
* Dane
  + Format danych
    - Ujednolicony format, dlaczego i po co
    - Jaki? Wybór nowego „zaprojektowanego przez nas”
    - Które wartości w danych są ważne i dlaczego, stąd wychodzi jaki format stworzyć
* Wczytywanie danych do programu
  + W jakiej postaci, wektor, tabela, lista, czemu, który
* Wyświetlanie danych
  + Wykresy
  + W każdym opisane co to za wykres, screen jak wygląda, czemu ma służyć, jakimi wartościami z danych się posługuje
    - Strona wyświetlania danych
    - Strzałkowy genom
    - Strzałkowy RNA-seq/microarray
    - Chip-seq
    - Tabele z danymi pod wykresami
    - Opisanie wszystkich przycisków, wyborów, zmiennych które można zastosować do modyfikacji wykresu
      * Zaprezentowanie screenami jak to działa, jakie są i jakie modyfikacje można wprowadzać na wykresie
        + Zmiana genomu
        + Wybór wyświetlanych wykresów
        + Wybór danych do wykresów
        + Wybór genu
        + Oddzielny wybór danych do chip-seq
        + Przyciski „przesuwania się” po wykresie
        + Granice wykresu
        + Granice logFC z tym jak się zmieniają
        + Przycisk filtra FDR
      * Dodanie, że jest opcja pobrania wykresu
    - Dodawanie własnych danych przez użytkownika
      * Opis jak dane należy dodawać, z dodaniem, że trzeba je sobie nazwać
      * Jakie manipulacje są wprowadzane w danych (przyrównanie do genomu z dodaniem start, end itp.) (trzeba gdzieś też będzie zamieścić wszystkie genomy których używaliśmy)
    - Strona porównania danych
    - Diagram venna, co pokazuje, jakie modyfikacje wprowadzane są do danych
    - Heat mapa
    - Wykres intime, jakie modyfikacje wprowadzane są do danych
    - Wszelkie przyciski modyfikacji wprowadzaych na wykresach
    - Opcja pobrania wykresów w formie obrazu
    - HELP$Info
      * Co jest w nim zawarte
        + Pobranie szablonu formatu danych
        + Może dodać do aplikacji żeby umiała pivot zrobić z danymi, to by było spoko, że jak ktoś wrzuci dane, to jeśli ktoś dobrze ponazywa kolumny, to żeby mogło wybrać kolumny, zrobić pivot, bo przestawianie danych sobie samemu w excelu jest problematyczne. Wiadomo że ultra uniwersalnie się tego nie zrobi ale może by działało
        + Trzeba tu też wstawić jakie dane są w aplikacji, taką bardziej rozbudowaną bibliografię
        + Podkreślić że jest tu opis tego co się dzieje w aplikacji, myślę, że zamiast opisywać wszystko w tym małym help&info, można będzie po prostu skopiować dużą część z sekcji materiały i metody która powstanie tutaj i wstawić jako plik do pobrania, taki pomysł mi do głowy przyszedł

# Wyniki

* Prezentacja kilku przykładów jak można zastosować aplikację pewnie też w wersji bardzo obrazkowej
  + Co mi pierwsze do głowy przychodzi, to:
    - Opcja, ktoś bazuje tylko na tym co jest w aplikacji,
      * Ogląda któreś zbiory które go zainteresowały na wykresach strzałkowych, przechodzi do zakładki comparison i tam ogląda dokładniej poszczególne geny/zbiory genów na wykresie i wyciąga jakieś wnioski, jakie? To się okaże, trzeba spróbować samemu coś tak pomyśleć, to znaczy ja

możliwe przykłady zastosowania:

analiza jednego białka np. chcemy się dowiedzieć w czy i w jakich warunkach zachodzi ekspresja genu, czy należy do konkretnego regulonu, operonu

analiza całego klastra genów np. chcemy sprawdzić w jakich warunkach aktywny jest klaster genów odpowiedzialnych za produkcję metabolitów wtórnych np. aktynorodyna lub red u S. coelicolor lub chloramfenikol u S. venezuelae

* + - Opcja, ktoś dodaje własne,
      * Podobnie jak z przykładem poniżej, tyle że wychodzi ktoś od swoich danych, znajduje, którąś publikacje, która ma podobne założenia co jego i ogląda co się dzieje w jego w porównaniu z tą drugą, może też sobie pobrać wykresy do zamieszczenia w publikacji

# Dyskusja/Podsumowanie

* Jak sobie poradziliśmy z problemami dotyczącymi danych
* Jakie korzyści może przynieść używanie bazy
* Jakie wyniki i wnioski można wyciągać
* Jak to pomoże ludziom
* Podkreślenie, że aplikacja jest w pełni uaktualnialna, to znaczy, będzie można dodawać nowe zbiory danych
* I pewnie też coś jeszcze co przyjdzie do głowy po napisaniu reszty
* Problematyka ujednolicania danych i jak to jest że porównujemy dane które zostały bazowo inaczej przygotowane
* Szczerze to nie wiem co tu jeszcze może być, pewnie więcej pomysłów przyjdzie do głowy na koniec

# Bibliografia

[1] E. A. Barka *et al.*, “Taxonomy, Physiology, and Natural Products of Actinobacteria,” *Microbiology and Molecular Biology Reviews*, vol. 80, no. 1, pp. 1–43, Mar. 2016, doi: 10.1128/MMBR.00019-15.

[2] S. D. Bentley *et al.*, “Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2),” *Nature*, vol. 417, no. 6885, pp. 141–147, May 2002, doi: 10.1038/417141a.

[3] M. Nikolaidis *et al.*, “A panoramic view of the genomic landscape of the genus Streptomyces,” *Microb Genom*, vol. 9, no. 6, Jun. 2023, doi: 10.1099/mgen.0.001028.

[4] S. E. Jones and M. A. Elliot, “‘Exploring’ the regulation of Streptomyces growth and development,” *Curr Opin Microbiol*, vol. 42, pp. 25–30, Apr. 2018, doi: 10.1016/j.mib.2017.09.009.

[5] G. A. Quinn, A. M. Banat, A. M. Abdelhameed, and I. M. Banat, “Streptomyces from traditional medicine: sources of new innovations in antibiotic discovery,” *J Med Microbiol*, vol. 69, no. 8, pp. 1040–1048, Aug. 2020, doi: 10.1099/jmm.0.001232.

[6] K. F. Chater, “Recent advances in understanding Streptomyces,” *F1000Res*, vol. 5, p. 2795, Nov. 2016, doi: 10.12688/f1000research.9534.1.

[7] D. L. Sexton and E. I. Tocheva, “Ultrastructure of Exospore Formation in Streptomyces Revealed by Cryo-Electron Tomography,” *Front Microbiol*, vol. 11, Sep. 2020, doi: 10.3389/fmicb.2020.581135.

[8] K. Alam *et al.*, “Streptomyces: The biofactory of secondary metabolites,” *Front Microbiol*, vol. 13, Sep. 2022, doi: 10.3389/fmicb.2022.968053.

[9] Z. Liu, Y. Zhao, C. Huang, and Y. Luo, “Recent Advances in Silent Gene Cluster Activation in Streptomyces,” *Front Bioeng Biotechnol*, vol. 9, Feb. 2021, doi: 10.3389/fbioe.2021.632230.

[10] S. Hwang *et al.*, “Streptomyces as Microbial Chassis for Heterologous Protein Expression,” *Front Bioeng Biotechnol*, vol. 9, Dec. 2021, doi: 10.3389/fbioe.2021.804295.

[11] S. Schlimpert and M. A. Elliot, “The Best of Both Worlds—Streptomyces coelicolor and Streptomyces venezuelae as Model Species for Studying Antibiotic Production and Bacterial Multicellular Development,” *J Bacteriol*, vol. 205, no. 7, Jul. 2023, doi: 10.1128/jb.00153-23.

[12] L. V Bystrykh, M. A. Fernández-Moreno, J. K. Herrema, F. Malpartida, D. A. Hopwood, and L. Dijkhuizen, “Production of actinorhodin-related ‘blue pigments’ by Streptomyces coelicolor A3(2),” *J Bacteriol*, vol. 178, no. 8, pp. 2238–2244, Apr. 1996, doi: 10.1128/jb.178.8.2238-2244.1996.

[13] D. Sun, C. Liu, J. Zhu, and W. Liu, “Connecting Metabolic Pathways: Sigma Factors in Streptomyces spp.,” *Front Microbiol*, vol. 8, Dec. 2017, doi: 10.3389/fmicb.2017.02546.

[14] E. Lee *et al.*, “A master regulator σ B governs osmotic and oxidative response as well as differentiation via a network of sigma factors in Streptomyces coelicolor,” *Mol Microbiol*, vol. 57, no. 5, pp. 1252–1264, Sep. 2005, doi: 10.1111/j.1365-2958.2005.04761.x.

[15] A. Romero-Rodríguez, I. Robledo-Casados, and S. Sánchez, “An overview on transcriptional regulators in Streptomyces,” *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1849, no. 8, pp. 1017–1039, Aug. 2015, doi: 10.1016/j.bbagrm.2015.06.007.

[16] M. J. Szafran, D. Jakimowicz, and M. A. Elliot, “Compaction and control—the role of chromosome-organizing proteins in Streptomyces,” *FEMS Microbiol Rev*, vol. 44, no. 6, pp. 725–739, Nov. 2020, doi: 10.1093/femsre/fuaa028.

[17] E. M. F. Shepherdson and M. A. Elliot, “Redefining development in Streptomyces venezuelae : integrating exploration into the classical sporulating life cycle,” *mBio*, vol. 15, no. 4, Apr. 2024, doi: 10.1128/mbio.02424-23.

[18] N. Tschowri, “Cyclic Dinucleotide-Controlled Regulatory Pathways in Streptomyces Species,” *J Bacteriol*, vol. 198, no. 1, pp. 47–54, Jan. 2016, doi: 10.1128/JB.00423-15.

[19] L. T. Fernández-Martínez *et al.*, “New Insights into Chloramphenicol Biosynthesis in Streptomyces venezuelae ATCC 10712,” *Antimicrob Agents Chemother*, vol. 58, no. 12, pp. 7441–7450, Dec. 2014, doi: 10.1128/AAC.04272-14.

[20] D. Muzzey, E. A. Evans, and C. Lieber, “Understanding the Basics of NGS: From Mechanism to Variant Calling,” *Curr Genet Med Rep*, vol. 3, no. 4, pp. 158–165, Dec. 2015, doi: 10.1007/s40142-015-0076-8.

[21] S. Marguerat and J. Bähler, “RNA-seq: from technology to biology,” *Cellular and Molecular Life Sciences*, vol. 67, no. 4, pp. 569–579, Feb. 2010, doi: 10.1007/s00018-009-0180-6.

[22] K. Van Den Berge *et al.*, “RNA sequencing data: hitchhiker’s guide to expression analysis,” Oct. 17, 2018. doi: 10.7287/peerj.preprints.27283v1.

[23] S. Antoraz *et al.*, “The Orphan Response Regulator Aor1 Is a New Relevant Piece in the Complex Puzzle of Streptomyces coelicolor Antibiotic Regulatory Network,” *Front Microbiol*, vol. 8, Dec. 2017, doi: 10.3389/fmicb.2017.02444.

[24] Y. Lu *et al.*, “An Orphan Histidine Kinase, OhkA, Regulates Both Secondary Metabolism and Morphological Differentiation in Streptomyces coelicolor,” *J Bacteriol*, vol. 193, no. 12, pp. 3020–3032, Jun. 2011, doi: 10.1128/JB.00017-11.

[25] Y. Jeong *et al.*, “The dynamic transcriptional and translational landscape of the model antibiotic producer Streptomyces coelicolor A3(2),” *Nat Commun*, vol. 7, no. 1, p. 11605, Jun. 2016, doi: 10.1038/ncomms11605.

[26] P. Salerno *et al.*, “Identification of new developmentally regulated genes involved in Streptomyces coelicolorsporulation,” *BMC Microbiol*, vol. 13, no. 1, p. 281, Dec. 2013, doi: 10.1186/1471-2180-13-281.

[27] P. Yagüe *et al.*, “Transcriptomic Analysis of Streptomyces coelicolor Differentiation in Solid Sporulating Cultures: First Compartmentalized and Second Multinucleated Mycelia Have Different and Distinctive Transcriptomes,” *PLoS One*, vol. 8, no. 3, p. e60665, Mar. 2013, doi: 10.1371/journal.pone.0060665.

[28] M. J. Szafran *et al.*, “Spatial rearrangement of the Streptomyces venezuelae linear chromosome during sporogenic development,” *Nat Commun*, vol. 12, no. 1, p. 5222, Sep. 2021, doi: 10.1038/s41467-021-25461-2.

[29] K. Šmídová, A. Ziková, J. Pospíšil, M. Schwarz, J. Bobek, and J. Vohradsky, “DNA mapping and kinetic modeling of the HrdB regulon in Streptomyces coelicolor,” *Nucleic Acids Res*, vol. 47, no. 2, pp. 621–633, Jan. 2019, doi: 10.1093/nar/gky1018.

[30] A. Shah, “Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiDTM system,” *Nat Methods*, vol. 6, no. 4, pp. ii–iii, Apr. 2009, doi: 10.1038/nmeth.f.247.

[31] U. Sarkans *et al.*, “The BioStudies database—one stop shop for all data supporting a life sciences study,” *Nucleic Acids Res*, vol. 46, no. D1, pp. D1266–D1270, Jan. 2018, doi: 10.1093/nar/gkx965.

[32] K. Rychel, K. Decker, A. V Sastry, P. V Phaneuf, S. Poudel, and B. O. Palsson, “iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning,” *Nucleic Acids Res*, vol. 49, no. D1, pp. D112–D120, Jan. 2021, doi: 10.1093/nar/gkaa810.

[33] S. X. Ge, E. W. Son, and R. Yao, “iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data,” *BMC Bioinformatics*, vol. 19, no. 1, p. 534, Dec. 2018, doi: 10.1186/s12859-018-2486-6.

# Skróty i symbole

Nawet nie wiem czy jakieś konkretne będą, ale raczej okaże się, że tak. Chociażby:  
R – język programowania.