Biologia Systemów 2024/25 - Spectrometria Mas i Spectroskopia NMR

Maria Bochenek
m.bochenek@uw.edu.pl

Termin oddania: ~~31 maja 2025 godzina 23:59~~ 11.06.2025 23:59.
Forma oddania: repozytorium na GitHub zawierające kod, figury oraz plik PDF z raportem (szczegóły poniżej)
Liczba punktów: max 10 punktów.

Wprowadzenie

W tym projekcie zajmiemy się analizą danych pochodzące z badania mającego na celu sprawdzenie jak zmienia się poziom metabolitów w moczu pochodzących od 6 wolontariuszy przebywających w ustandaryzowanym i stabilnym środowisku. Wszyscy wolontariusze dzielą ten sam rytm dobowy, dietę oraz są wystawieni na działanie tych samych czynników zewnętrznych.

Hipoteza badawcza zakłada, że im dłużej wolontariusze będą przebywać w tym samym środowisku tym bardziej podobny będzie skład metabolitów w ich moczu. Oznaczałoby to, że to środowisko zewnętrzne ma większy wpływ na nasz metabolizm niż różnice osobnicze. W celu zweryfikowania tej hipotezy badacze dokonali analizy próbek moczu przy pomocy spektroskopii 1H NMR. Próbki były zbieranie codziennie na przestrzeni 254 dni trwania eksperymentu oraz w dniach 713-716 od początku eksperymentu. Dodatkowo zmieszano ze sobą wszystkie próbki zebrane podczas eksperymentu oraz dokonano 32 pomiarów tak stworzonej próbki kontrolnej (QC).

W naszych analizach skupimy się na analizie 29 widm NMR próbek pochodzących od jednego wolontariusza oraz jednego widma próbki QC z powodu ograniczeń w zasobach obliczeniowych. Widma NMR znajdują się w plikach csv postaci `X_V5001_DY.csv`, gdzie `X` to unikatowy numer próbki a `Y` to numer dnia, w którym została pobrana próbka moczu. Widmo QC znajduje się w pliku `1541_QC.csv`. W pierwszej kolumnie w plikach znajdują się pozycje pików podane w ppm a w drugiej kolumnie znajdują się intensywności pików.

Zasady oceniania

Celem projektu jest wykonanie analizy widm NMR i przygotowanie raportu z wykonanych analiz. Raporty mogą być pisane po polsku lub po angielsku.

W notebooku `BS_24/25_NMR_student_version.ipynb` znajduje się sekcja "HMDB and urine sample analysis", w której opisane są szczegółowo zadania obliczeniowe. Znajdują się w niej również pytania pomocnicze, które mogą być przydatne przy pisaniu raportu.

W przypadku problemów wynikających z ograniczonych zasobów obliczeniowych można ograniczyć liczbę analizowanych próbek. Jakiekolwiek modyfikacje muszą być opisane w raporcie.

Oceniana będzie jakość raportu, jasność przekazu i sposób prezentacji wyników. Punktacja za raport podana jest przy poszczególnych sekcjach. Raport powinien składać się z sekcji opisanych poniżej.

Raport (3 pkt)

1. (0.25 pkt) Wprowadzenie - napisz krótki wstęp teoretyczny uwzględnij w nim informacje takie jak
   - Czym jest spektroskopia NMR?
   - Czym jest metabolomika?
   - Jakie zastosowanie ma spektroskopia NMR w metabolomice?
   - Jaki jest cel przeprowadzonych analiz (co robimy i po co)?
2. Metody
   - (0.5 pkt) Opisz zbiór danych eksperymentalnych. Wyjaśnij jak widma eksperymentalne zostały przygotowane do analiz (preprocessing).
   - (0.5 pkt) Opisz metodologię przygotowania bibliotek widm referencyjnych. To jest dobre miejsce na odpowiedzi do zadania 1 i zadania 3.1.
   - (0.25 pkt) Opisz wybór parametrów κ i solverów. to jest dobre miejsce na odpowiedzi do zadania 2, ale uwaga porównania i wizualizacje z zadania 2.2 i 2.4 umieść w sekcji wyniki.
3. (1.25 pkt) Wyniki
   - Wykresy powinny być opisane oraz w tekście powinno się znaleźć do nich nawiązanie. Nie umieszczaj wykresów, których nie komentujesz w tekście. Jeżeli chcesz to zrobić umieść je w sekcji "Dodatkowe materiały".
   - Uwzględnij tutaj wyniki z zadania 2 oraz zadania 3.2.
4. (0.25 pkt) Dyskusja
5. Bibliografia

Zadanie 1 (1.75 pkt)

1. Czym jest HMDB i jakie dane możemy w niej znaleźć?
2. Pobierz wszystkie widma NMR zawarte w bazie HMDB.
3. (1 pkt) Napisz funkcję `extract_1D_spectra(folder_path: str, nucleus: str)`, która jako argument przyjmuje zmienną `folder_path` kierującą do folderu z plikami xml zawierającymi widma oraz `nuclei` - rodzaj jądra atomowego (np. 1H, 13C). Funkcja ma wyszukiwać w pliku xml przesunięcie chemiczne (chemical shift), pozycję pików (peak

position) w jednostkach ppm oraz intensywność dla każdego z widm 1D i zapisywać je do pliku csv o nazwie postaci:

`{HMBD_ID}_1D_{Spectrum_ID}_{pred/exper}_{nucleus}_{frequency}.csv`.

Przykład 1 Zawartość pliku `HMDB0000064_1D_142013_predicted_1H_1000.csv`:

```
chemical_shift,peak_position_ppm,intensity
3.02,3.020034790039063,0.0020594516
3.02,3.020034790039063,0.001297893
3.02,3.020034790039063,0.0017210965
3.92,3.920028686523438,0.0016928137
3.92,3.920028686523438,0.0016928137
```

Przykład 2 Zawartość pliku `HMDB0000064_1D_1064_experimental_1H_500.csv`:

```
chemical_shift,peak_position_ppm,intensity
3.92,3.92,55.03
3.02,3.02,100.0
```

4. (0.5 pkt) Napisz funkcję `preprocess_1D_spectra(filename:str, out_folder="./Preprocessed", sig=None)`, która wczytuje plik w formacie stworzonym przez funkcję `extract_1D_spectra` a następnie
   (a) zaokrągla (to ma być argument opcjonalny) pozycje pików do sig liczb znaczących
   (b) grupuje piki po ich pozycji i przypisuje im nową intensywność równą sumie intensywności wszystkich pików w grupie. (*Podpowiedź*: `pandas.DataFrame.groupby()`, `sum()`, `reset_index()`.)
   (c) zapisuje przetworzone widma w folderze `out_folder` (tworzy go jeśli taki folder nie istnieje) w plikach o nazwie `filename_processed.csv`
5. Jak powinniśmy przetworzyć pobrane widma zanim użyjemy ich jako widma referencyjne do estymacji proporcji metabolitów w mieszaninie? Wyjaśnij dlaczego wykonujemy te operacje.
6. Ile metabolitów z HMDB wykrytych w moczu (detected) posiada eksperymental widma 1D 1H NMR? Ile wykrytych i skwantyfikowanych (detected and quantified)? Ile z wykrytych i nieskwantyfikowanych?
7. Ile metabolitów z pliku `selected_metabolites.csv` posiada widmo eksperymentalne. Jakie są najwyższe dostępne częstotliwości widm symulowanych dla tych metabolitów?
8. (0.25 pkt) Wybierz po jednym spektrum dla każdego metabolitu. Uzasadnij swój wybór.

Zadanie 2 (3.75 pkt)

1. (1 pkt) Napisz funkcję, która rysuje wyniki regresji widm (rysuje modelowe widmo) oraz widmo eksperymentalne. Tutaj dobrym pomysłem może być wizualizowanie mniejszego fragmentu widm.

2. (0.25 pkt) Oszacuj proporcje wybranych metabolitów dla kilku widm eksperymentalnych oraz kontroli. Porównaj wyniki z użyciem dwóch różnych solverów (przydatna będzie funkcja z powyższego punktu). Opisz różnice. Uzasadnij jakiego solvera będziesz używać w analizach.

3. (0.25 pkt) Wyjaśnij czym są parametry κ w używanym modelu.

4. (0,25 pkt) Porównaj wyniki dla kilku wartości parametrów κ. Czy udało ci się znaleźć parametry, które dają lepsze wyniki niż domyślne wartości? Wybierz najlepsze parametry do dalszej analizy. Uzasadnij swój wybór.

5. (1 pkt) Używając wybranego solvera i wartości κ oszacuj proporcje wybranych metabolitów w próbkach eksperymentalnych i kontroli.

6. (1 pkt) Jak zmieniają się proporcje metabolitów w zależności od czasu eksperymentu (dni)? Przygotuj odpowiednie wizualizacje. Opisz wyniki.

Zadanie 3 (1.5 pkt)

1. (0.5 pkt) Skonstruuj nową bibliotekę widm referencyjnych, które można znaleźć w moczu. Możesz

   - stworzyć całkiem nową bibliotekę referencyjną (nie może być mniejsza niż od biblioteki z wcześniejszych zadań),
   - rozszerzyć używaną wcześniej bibliotekę o kilka metabolitów,
   - uwzględnić tylko część metabolitów z poprzedniej biblioteki i rozszerzyć ten zbióro nowe metabolity.

   Uzasadnij wybraną strategię. Opisz nową bibliotekę referencyjną. Jakie nowe metabolity zostały uwzględnione i dlaczego?

2. (1 pkt) Oszacuj proporcje metabolitów w próbkach przy użyciu nowej biblioteki referencyjnej dla kilku widm eksperymentalnych oraz kontroli. Porównaj wyniki z wynikami z zadania 3. Przygotuj odpowiednie wizualizacje.

Literatura

[1] Majewski, S.; Ciach, M. A.; Startek, M.; Niemyska, W.; Miasojedow, B.; Gambin, A. *The Wasserstein Distance as a Dissimilarity Measure for Mass Spectra with Application to Spectral Deconvolution.* 18th International Workshop on Algorithms in Bioinformatics (WABI 2018). 2018.
[2] Ciach, M. A.; Miasojedow, B.; Skoraczyński, G.; Majewski, S.; Startek, M.; Valkenborg, D.; Gambin, A. *Masserstein: Linear regression of mass spectra by optimal transport.* Rapid Communications in Mass Spectrometry 2020, e8956.

[3] Domżał, B.; Nawrocka, E. K.; Gołowicz, D.; Ciach, M. A.; Miasojedow, B.; Kazimierczuk, K.; Gambin, A. *Magnetstein: An Open-Source Tool for Quantitative NMR Mixture Analysis Robust to Low Resolution, Distorted Lineshapes, and Peak Shifts.* Analytical Chemistry 2023, 96, 188–196.

Forma oddania

- Kod, figury oraz raport powinny znajdować się w repozytorium Github (można np zrobić fork repozytorium `magnetstein` i w nim pracować).
- Analizy mogą być wykonane z użyciem skryptów pythonowych lub Jupyter Notebook'ów.
- Kod powinien być reprodukowalny oraz jasno skomentowany.
- Skrypty/notebooki z waszym kodem mają być łatwe do zlokalizowania w repozytorium.
- W repozytorium powinien znajdować się folder z figurami uwzględnionymi w raporcie.
- Raport ma być samowystarczalny do zrozumienia waszych analiz, bez potrzeby zaglądania do kodu.
- Gdy wasze repozytorium będzie gotowe dodajcie mnie (mariaboch) oraz prowadzącego jako kolaboratorów oraz uzupełnijcie formatkę na Moodle.

Kontakt

W razie jakichkolwiek wątpliwości co do poleceń proszę o kontakt `m.bochenek@uw.edu.pl`.