Genomika dla informatyków – zadania kwalifikacyjne

Jako rozwiązanie należy przesłać kod źródłowy oraz narysowane wykresy. Termin oddania: 11.07.2017, mailem na adres `ajank(at)mimuw.edu.pl`. Wszelkie pytania należy kierować na ten sam adres. :)

Jak wiadomo, ekspresja genów przebiega dwuetapowo: wpierw w procesie transkrypcji sekwencja DNA jest odczytywana i powstaje matrycowe RNA, po czym następuje translacja, czyli synteza białka na podstawie matrycowego RNA. Zajmiemy się danymi z eksperymentów sekwencjonowania transkryptomu (RNA-seq), który dokonują pomiaru na tym pierwszym etapie. Eksperymenty były wykonane na linii komórkowej A549, uzyskanej z ludzkich komórek raka płuc (gruczolakoraka) pobranych od pacjenta.

W części eksperymentów, komórki zostały poddane przez 12 godzin działaniu deksametazonu. Deksametazon jest lekiem o działaniu przeciwzapalnym, przeciwalergicznym i immunosupresyjnym, podawanym m.in. wspomagająco przy chemioterapii. Naszym celem będzie porównanie ekspresji genów tych w dwóch warunkach eksperymentalnych, oraz ustalenie, które geny mają najbardziej zmienioną aktywność.

Część I – techniczna

Korzystamy z publicznie dostępnych danych, uzyskanych przez zespół Tima Reddy'ego z Duke University. Pod adresem `https://www.mimuw.edu.pl/~ajank/WWW13/` znajduje się podzbiór tych danych, ograniczony do chromosomów 20 i 21. W katalogach `0h` i `12h` znajdują się pliki w formacie BED ze współrzędnymi genomowymi odczytów z eksperymentów RNA-seq. W każdym z tych dwóch warunków eksperymentalnych wykonano cztery niezależne eksperymenty. Ponadto plik `genes_hg38.bed` zawiera współrzędne genów na tych dwóch chromosomach.

1. Dla każdego genu, policzyć ile odczytów RNA-seq zachodzi na niego (tzn. dwa przedziały mają niepustą część wspólną) w każdym z eksperymentów. Istotne jest, aby brać pod uwagę tylko odczyty z tej samej nici DNA (`+` lub `-`) co gen.
2. Dla każdego genu, policzyć sumę i odchylenie standardowe ww. liczb w obu warunkach eksperymentalnych (Control – bez działania deksametazonu, Treatment – po działaniu deksametazonu).

Aby to łatwo zrobić, można wykorzystać np. pakiet `GenomicRanges` w R, pakiet `pybedtools` w Pythonie lub zestaw narzędzi `BEDTools`.

Część II – eksploracyjna

3. Narysować wykres, na którym geny zaznaczone są punktami, zaś współrzędne odpowiadają sumie odczytów RNA-seq: na osi x Control, na osi y Treatment.
4. Narysować tzw. MA plot przedstawiający te same dane w inny sposób: na osi x Mean Average $= \frac{1}{2}\log_2(\text{Treatment} \cdot \text{Control})$, na osi y Log Ratio $= \log_2(\text{Treatment}/\text{Control})$.
5. Zaproponować kryterium wskazujące geny o najbardziej zmienionej aktywności wskutek działania deksametazonu. Wybrać pięć genów o najbardziej zmienionej aktywności, oraz zaznaczyć je innym kolorem na wykresach. Można zobaczyć, co to za geny, wyszukując je po identyfikatorze w bazie danych Ensembl.

Oczywiście można korzystać z dowolnych narzędzi, np. standardowej biblioteki graficznej lub `ggplot2` w R, albo `matplotlib` w Pythonie.