Silesian University of Technology

Faculty of Automatic Control, Electronics and Computer Science

Advanced data exploration techniques for augmented transcriptional landscape and its better quantification

PhD Thesis

Author: Agata Muszyńska
Supervisor: dr hab. inż. Paweł Łabaj
Co-supervisor: Assoc. Prof. Dr. David Kreil

Gliwice, February 2023

Zaawansowane techniki analizy danych ekspresyjnych dla lepszego zrozumienia transkryptomu i poprawy jego oceny ilościowej

Wydział Automatyki, Elektroniki i Informatyki
Politechnika Śląska

Streszczenie rozprawy doktorskiej
Autor: Agata Muszyńska
Promotor: dr hab. inż. Paweł Łabaj
Kopromotor: Assoc. Prof. Dr. David Kreil

W ostatnich latach badacze stają się coraz bardziej świadomi kryzysu powtarzalności, z jakim boryka się całe środowisko naukowe. Pomimo dostarczonych kompleksowych testów porównawczych i wytycznych, problem ten jest również ważnym zagadnieniem w przypadku analizy danych RNA-seq. Powodem tego jest bardzo często brak stosowania tych wskazówek w praktyce, a także korzystanie z narzędzi, które są przestarzałe, ale po prostu dobrze znane. Innym, bardziej złożonym powodem braku powtarzalności wyników sekwencjonowania może być złożoność transkryptomu. Chociaż ludzki genom jest najbardziej zbadany i kompletny, wciąż pojawiają się nowe aktualizacje referencji, różniące się liczbą genów i transkryptów.

Celem tej pracy jest wybranie najlepszych praktyk w analizie danych RNA-seq i opracowanie podejścia prowadzącego do poprawy powtarzalności i stabilności wyników. Przedstawione tutaj rozwiązanie opiera się na istniejących badaniach i dostępnych narzędziach, ale modyfikuje je i łączy w dostosowany system przetwarzania potokowego. Wyjątkowość tego rozwiązania polega na tym, że integruje ono wiele etapów analizy danych RNA-seq, w przeciwieństwie do istniejących podejść, które koncentrują się na poszczególnych modułach. Ponieważ w metodach analizy danych o wysokiej przepustowości nie istnieją złote standardy, zapewnione są również różne opcje dla niektórych etapów. Ponadto zawarte są narzędzia do analizy alternatywnego splicingu (AS) (wykrywanie AS i analiza jego konsekwencji na poziomie transkryptu i białka) oraz wizualizacji. Tej części brakuje w standardowych podejściach, ma ona jednak kluczowe znaczenie dla lepszego zrozumienia i opisywania modeli genów. Przedstawione tutaj rozwiązanie wypełnia tę lukę, zapewniając kompleksowe narzędzie - od surowych odczytów, poprzez dopasowanie do referencji, analizę AS i genów różnicujących, aż po wyniki na poziomie białkowym.

Ta praca przedstawia wyniki dla trzech różnych zestawów danych - „rzeczywisty" przykład danych next-generation sequencing (NGS), benchmarkingowe dane NGS i dane mikromacierzowe. Każdy zestaw danych dostarczył nowych spostrzeżeń i ulepszeń w budowie narzędzia, a jednocześnie potwierdził wybór metod. Wszystkie podejścia potwierdziły, że nie ma uniwersalnego rozwiązania zarówno w zakresie metody analizy danych, jak i doboru technik laboratoryjnych. Dlatego bardzo korzystne jest zapewnienie opcji w narzędziach służących analizie danych o dużej przepustowości, tak aby można było wybrać ścieżkę najlepiej pasującą do celu eksperymentu. Jest to szczególnie ważne w przypadku rzeczywistych scenariuszy, w których sygnał może być słaby, a prawdziwe zaburzenia danych mogą być nieznane.

Ta praca pokazała również, że złożoność genomu myszy, ale także ludzkiego, nie jest jeszcze w pełni poznana. Pomimo obecności wielu czynników zakłócających sygnał, udało się wykryć wspólne wzorce w danych. Oczywiste było, że wspólne terminy ontologii genów dla różnych typów alternatywnego splicingu były związane z układem nerwowym. Wydaje się, że funkcje i procesy, w które są zaangażowane, są charakterystyczne dla danego typu AS. Wyniki wskazują, że istnieje wiele

wariantów, które nie są jeszcze obecne w referencji i że nadal można wprowadzić znaczne ulepszenia w dziedzinie referencji transkryptomu. Ta obserwacja jest zgodna z kilkoma niedawnymi artykułami opisującymi wiele nowych wariantów AS, występujących w różnych obszarach mózgu, a także w innych tkankach u różnych gatunków. Mimo to, wciąż brakuje kompleksowego narzędzia do badania tego aspektu. W wyniku tych badań opracowano system przetwarzania potokowego obejmujący różne aspekty analizy danych o dużej przepustowości, wypełniający tę lukę. Pomoże to poprawić powtarzalność, zapewniając większą kontrolę nad metodami i parametrami stosowanymi na różnych etapach analizy, ale także zaspokajając potrzebę automatycznego wykrywania i analizy AS oraz jego konsekwencji. Ponieważ AS odgrywa również ogromną rolę w wielu chorobach, narzędzie to może dostarczyć nowych informacji na temat zmian zachodzących w stanach takich jak choroba Parkinsona, SMA lub różne typy raka.