Metody Bioinformatyki

Wykorzystanie PCA do analizy danych z mikromacierzy DNA

Maciej Czerniak, Marcin Kamionowski, Kacper Szkudlarek

29 listopada 2011

Streszczenie

Dokumentacja realizacji projektu z przedmiotu "Metody Bioinformatyki". W ramach projektu wykonane zostanie oprogramowanie wykorzystujące metodę Analizy Składowych Głównych (PCA) do przetwarzania danych uzyskanych z mikromacierzy DNA.

1 Wstęp

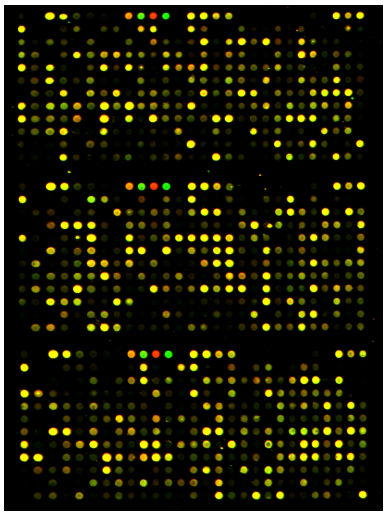
Analiza głównych składowych (ang. Principal Component Analysis, PCA) jest jedną ze statystycznych metod analizy czynnikowej. Zbiór danych składający się z N obserwacji, z których każda obejmuje K zmiennych, można interpretować jako chmurę N punktów w przestrzeni K-wymiarowej. Celem PCA jest taki obrót układu współrzędnych, aby maksymalizować w pierwszej kolejności wariancję pierwszej współrzędnej, następnie wariancję drugiej współrzędnej, itd.. Tak przekształcone wartości współrzędnych nazywane są ładunkami wygenerowanych czynników (składowych głównych). W ten sposób konstruowana jest nowa przestrzeń obserwacji, w której najwięcej zmienności wyjaśniają początkowe czynniki. PCA jest często używana do zmniejszania rozmiaru zbioru danych statystycznych, poprzez odrzucenie ostatnich czynników.

Mikromacierz DNA jest to płytka szklana lub plastikowa z naniesionymi w regularnych pozycjach mikroskopowej wielkości polami (ang. spots), zawierającymi różniące się od siebie sekwencją fragmenty DNA. Fragmenty te są sondami, które wykrywają przez hybrydyzację komplementarne do siebie cząsteczki DNA lub RNA.

Dane (Rys: 1) uzyskiwane w eksperymentach prowadzonych z wykorzystaniem mikromacierzy to wartości intensywności czerwonej oraz zielonej fluorescencji każdego z pół na płytce. Jedonorazowo w eksperymencie możliwe jest badanie ekspresji kilku tysięcy genów, dlatego uzyskane dane są wysoce złożone i wielowymiarowe.

2 Implementacja

Implementacje podzielona będzie na kilka części:



Rysunek 1: Pokolorowana próbka danych z mikromacierzy cDNA

1. Implementacja modułu pozwalającego na wczytywanie danych dostarczonych w surowej postaci plików tekstowych (tabbed separate data). Plik będzie analizowany pod kątem spójności i zgodności z formatem danych. Dane będą zapisywane w pamięci sposób pozwalający na ich dalszą obróbkę i analizę.

2. Stworzenie graficznego interfejsu użytkownika. Prosty interfejs pozwalający na wczytywanie danych, uruchomienie/zatrzymanie analizy, a także przejście do wizualizacji danych przed i po przetworzeniu oraz obliczenie błędu średniokwadratowego. W implementacji zostanie wykorzystana mulitplatformowa biblioteka Qt.

3. Moduł analizatory i przetwarzający dane wejściowe. Analiza danych będzie polegała na poddaniu ich analizie składowych głównych (PCA). Wykorzystana zostanie implementacja PCA znajdująca się w bibliotece OpenCV.

Całość implementacji zostanie wykonana w języku C/C++ pod kontrolą systemu Linux. Dzięki użyciu multiplatformowych bibliotek aplikacja będzie w pełni przenośna, wymagać będzie jedynie rekompilacji pod kontrolą systemu docelowego.