Elementy Bioinformatyki – Projekt 2

Maciej Ciszewski	119297
Kajetan Targoński	119483
Maciej Tyszka	119488
Dawid Zima	119509

Spis treści

Temat

Rozpoznanie możliwości bibliotek bioinformatycznych dla platformy .NET

Wybór biblioteki

Istnieje bardzo wiele bibliotek bioinformatycznych dla różnych platform i języków programowania: BioPython, BioJava itp. Dla platformy .NET udało się odnaleźć jedną pozycję: .NET Bio. Projekt powstał jako Microsoft Biology Fundation w laboratoriach firmy Microsoft, a następnie został zarzucony i udostępniony szerokiemu gronu odbiorców i developerów jako open source na licencji Apache 2.0.
Projekt i dokumentacja są dostępne pod adresem: http://bio.codeplex.com/

Instalacja

Istnieją dwie możliwości rozpoczęcia pracy z biblioteką. Pierwszą jest instalacja w systemie po pobraniu pliku instalacyjnego. Drugą, pobranie kodu źródłowego ze strony Codeplex i dołączenie go do naszego projektu.

Wykorzystanie

W ramach projektu stworzona została aplikacja konsolowa w języku C#, prezentująca kilka typowych zastosowań bibliotek bioinformatycznych:

1. Wyszukiwanie różnic pomiędzy dwiema sekwencjami białkowymi.
2. Konkatenacja sekwencji proteinowych.
3. Usuwanie z sekwencji proteinowej znaków niezgodnych z jej alfabetem.
4. Konwersja jednego formatu sekwencji do innego.

Według dokumentacji, możliwych zastosowań jest o wiele więcej. Dostępne są m. in.:

- natywne wsparcie dla web service'ów BLAST oraz ClustalW – nie udało się uruchomić
- wsparcie dla przetwarzania równoległego przy pomocy technologii HPC (High Performance Computing)
- zaimplementowana reprezentacja dla drzew filogenetycznych
- implementacja przydatnych algorytmów takich jak: Boyera-Moora do poszukiwania wzorców, budowania drzew sufiksowych czy wyrównywania sekwencji

Biblioteka została tak zaprojektowana, aby użytkownik mógł bardzo łatwo rozszerzyć jej możliwości według własnych potrzeb. Zostały zdefiniowane szczegółowe interfejsy, np.

parserów czy formaterów sekwencji, dzięki czemu dodanie własnych klas nie stanowi dużego wyzwania.

Wady i zalety

Niewątpliwą zaletą biblioteki .NET Bio jest jej wysoka elastyczność, dzięki której możliwe jest rozszerzanie możliwości przez użytkownika. Dodatkowym atutem jest pełen dostęp do kodu źródłowego oraz gotowe implementacje podstawowych operacji na drzewach, sekwencjach czy zbiorach.

Największą wadą okazuje się brak dobrze zredagowanej dokumentacji (nawet 'z kodu') oraz brak bardziej zaawansowanych przykładów zastosowań. Przez takie podejście, nie udało się w ramach projektu zbadać m. in. możliwości wykorzystania predefiniowanych web service'ów – stało by się to zapewne możliwe dopiero po bliższym zapoznaniu się z kodem źródłowym tych funkcjonalności, jednak nie powinno to być wymogiem przy korzystaniu z zewnętrznych, gotowych rozwiązań.

Poważnym mankamentem jest też duża ilość błędów, które oczekują na naprawienie. Niemożliwe np. okazało się konwertowanie sekwencji z jednego formatu w drugi czy poprawne wykrywanie różnic w sekwencjach, pomimo dostępnego przykładowego rozwiązania prezentowanego na stronie domowej.

Podsumowanie

Biblioteka .NET Bio posiada bardzo duży potencjał rozwojowy ze względu na ciekawą, intuicyjną i łatwo rozszerzalną architekturę i zastosowane zaawansowane technologie, takie jak HPC czy WCF w implementacji web service'ów.

Niestety, wydaje się, że dopóki nie zostaną naprawione poważne błędy oraz nie powstanie lepszej jakości, czytelniejsza dokumentacja, większym powodzeniem będą cieszyły się narzędzia o ugruntowanej pozycji takie jak: BioJava czy BioPython.

Linki

- MS Biology tools: http://research.microsoft.com/en-us/projects/bio/mbt.aspx
- Strona domowa projektu: http://bio.codeplex.com/
- Kod źródłowy projektu wykonanego w ramach przedmiotu: https://github.com/maciektys/bioInf