Nr rejestracyjny: 2017/27/B/ST6/02720; Kierownik projektu: dr hab. Paweł Piotr Górecki

Koherentne Modele i Wydajne Algorytmy dla Duplikacji Genomowych

Streszczenie

Głównym celem projektu jest opracowanie nowych biologicznie znaczących modeli i efektywnych algorytmów do rekonstrukcji duplikacji genomowych. W celu realizacji zadań tego interdyscyplinarnego projektu powołamy międzynarodowy zespół ekspertów złożony z biologów, matematyków i informatyków. Cele naukowe definiujemy poniżej:

1. Opracowanie biologicznie znaczących modeli dla problemu duplikacji genomowych.
2. Wyspecyfikowanie praktycznych problemów bazujących na tych modelach.
3. Analiza złożoności obliczeniowej tych problemów i zaprojektowanie efektywnych algorytmów dla nich.
4. Wydajna implementacja opracowanych algorytmów z przyjaznym dla biologów interfejsem.
5. Ewaluacja stosowalności i skalowalności opracowanych narzędzi.
6. Zastosowanie tych nowych narzędzi do rekonstrukcji i badania zjawisk duplikacji genomowych dla dużych zbiorów z rzeczywistymi danymi.

Tematyka zaproponowana w tym projekcie jest w kręgu zainteresowań nie tylko badaczy z zakresu biologii obliczeniowej i bioinformatyki, ale także dotyczy ważnych biologicznych i praktycznych zastosowań. W szczególności, duplikacje genomowe i cało-genomowe występowały w wielu gatunkach. Na przykład, istotnie kształtowały ewolucję roślin zbożowych, tak szczególnie ważnych dla rolnictwa i przemysłu.

Będziemy częściowo korzystać z naszych ostatnio opublikowanych wyników. Pierwszy to najlepszy obecnie dostępny algorytm, który umożliwia szybkie rozwiązywanie kilku wariantów problemu duplikacji genomowych. Ponadto, dzięki niemu rozwiązaliśmy bardziej ogólny, skomplikowany i jednocześnie praktyczny problem dla nieukorzenionych drzew genów. Jednakże, te wyniki doprowadziły także do powstania nowych pytań dotyczących zastosowanych modeli i jakości osiągniętych rekonstrukcji. Co więcej, wciąż jest wiele problemów otwartych w tej dziedzinie. Zastosowujemy te dwa rozwiązania jako punkty startowe by opracować nowe biologicznie znaczące modele i wydajne algorytmy dla problemu duplikacji genomowych.

Metody, które będziemy stosować to techniki teorii grafów, standardowe metody dowodzenia, algorytmy regułowe, optymalizacyjne typu “hill-climbing”, FPT, aproksymacyjne oraz genetyczne.

Narzędzia będą zaimplementowane w językach programowania C++ i Python. Przygotowanie danych i walidacja wyników, będzie wykonana przy użyciu specjalnie zaprojektowanych pipeline’ów. W takich jednorodnych środowiskach przygotujemy zbiór empirycznych i sztucznie wygenerowanych zestawów danych do testów. Walidacja będzie przeprowadzona przy ścisłej współpracy z ekspertami biologicznymi. Zakładamy, że walidacja dostarczy istotnych informacji dla zadań związanych z poszukiwaniem modeli i algorytmów.