

# BioSeq

## 1. Podstawowa obsługa aplikacji.

Aplikacja BioSeq pozwala na zestawienie dwóch sekwencji nukleotydowych, wyszukując te najbardziej podobne. Aplikacja pozwala zdefiniować kilka dodatkowych parametrów, takich jak macierz podobieństwa lub dystansu oraz składanie wykorzystujące kody genetyczne aminokwasów. W ten sposób jesteśmy w stanie manipulować prawdopodobieństwami zależnie od uznania, wyszukując tych najbardziej zbliżonych prawdzie. Aplikacja sprawdza się, gdy posiadamy dwie sekwencje nukleotydowe (w formie RNA, DNA lub aminokwasowej), które są ze sobą spokrewnione i próbujemy odtworzyć najbardziej prawdopodobne zmiany, które przebiegły w sekwencjach.

## 2. Dane wejściowe, wyjściowe

Na wejście aplikacja przyjmuje kody DNA (RNA) w formie sekwencji jako pliki tekstowy. Należy podać ścieżki do plików tekstowych z sekwencjami jako argumenty

`"--seq_A_path"` - dla pierwszego pliku z sekwencją

`"--seq_B_path"` - dla drugiego pliku z sekwencją

Oprócz tego wymagana jest plik, który pozwoli aplikacji na translację aminokwasową

`"--aminoacid_dict_path"` - ścieżka do pliku z translacją trójek nukleotydowych na aminokwasy

Oprócz tego możemy wykorzystać inne argumenty w celu manipulacji i poprawy wyników:

`"--convert_2_amino"` - flaga (domyślnie *false*) która ustawia czy plik ma wykorzystywać translację trójek nukleotydowych na aminokwasy w trakcie zestawiania.

`"--convert_2_dna"` - flaga (domyślnie *false*) która ustawia, czy aplikacja powinna w trakcie działania wykonać translację z RNA na DNA (ustawić na *true* jeśli na wejście podajemy sekwencję RNA)

`"--aligner_mode"` - czy zestawiania powinny być globalne (domyślnie) czy lokalne. Lokalne mogą być przydatne w celu znalezienia pomniejszych sekwencji, które mogą być podobne w podanych plikach.

`"--distance_matrix"` - ścieżka do pliku csv z macierzą dystansów, która pozwala określić koszt poszczególnych zmian nukleotydów na inne.

Na wyjściu aplikacja wyświetla najbardziej trafne zestawienie oraz uzyskaną ilość „punktów” przez dane zestawienie, które czynią je najbardziej prawdopodobne według podanych argumentów.

### 3. Algorytmy/biblioteki zastosowane

Aplikacja bazuje w dużej mierze na bibliotece Biopython, stworzoną przez grupę deweloperów dla programowania biologii molekularnej, co zawiera również zestawienia sekwencji nukleotydowych. Oprócz tego wykorzystywane są standardowe biblioteki Pyhotnowe do manipulacji danych/argumentów. Aplikacja sama w sobie nie posiada żadnych skomplikowanych algorytmów, wszystkie obciążające obliczenia wykonywane są w ramach biblioteki Biopython.

### 4. Możliwości biblioteki, klasy

Biblioteka Biopython udostępnia przeróżne możliwości przydatne przy manipulacji i analizie molekularnej w zakresie bioinformatyki, wszystko w granicach języka Python. Wśród nich możemy wyróżnić:

- Obsługa takich formatów jak:
  - o Blast jako wyjściowy (zarówno zwykły jak i www)
  - o Clustalw
  - o FASTA
  - o GenBank
  - o PubMed i Medline
  - o pliki ExPASy, takie jak Enzyme czy Prosite
  - o SCOP (w tym „dom” i „lin”)
  - o UniGene
  - o SwissProt
- Na wszystkich powyższych formatach można iterować po kolei lub za pomocą interfejsu słownikowego. Kod zawarty w Biopythonie obsługuje również popularne serwisy bioinformatyczne takie jak NCBI oraz ExPASy. Interfejsy do popularnych programów bioinformatycznych, takich jak:
  - Standardowy Blast z NCBI
  - Program do zestawień Clustalw
  - Zestaw operacji linii komend EMBOSS
- Podstawowe klasy do obsługi sekwencji, ids na sekwencjach oraz cech sekwencji
- Narzędzia do wykonywania typowych operacji na sekwencjach, takich jak translacje, transkrypcje, obliczenia na wagach
- Kod do wykonywania klasyfikacji danych metodą k najbliższych sąsiadów, Naiwnego Bayesa lub Maszyny Wektorów Nośnych

- Kod do obsługi zestawień, w tym standardowy sposób do tworzenia i obsługi macierzy podstawień
- Kod pozwalający na łatwe podzielenie zadań na osobne procesy
- Programy GUI do wykonywania dostawowych manipulacji na sekwencjach, translacji, BLASTowaniu itp.
- Bogata dokumentacja i pomoc przy korzystaniu z poszczególnych modułów.
- Integracja z BioSQL, schema bazy danych sekwencji wspieranych przez BioPerl oraz projekty BioJava.

## 5. Linki

Główna strona Biopythona: <https://biopython.org/>

Tutorial/ogólna

<http://biopython.org/DIST/docs/tutorial/Tutorial.pdf>

Szczegółowa dokumentacja: <http://biopython.org/DIST/docs/api/>

dokumentacja: