**Nazwa przedmiotu:**

Metody bioinformatyki

**Koordynator przedmiotu:**

Rajmund Kożuszek

**Status przedmiotu:**

Fakultatywny dowolnego wyboru

**Poziom kształcenia:**

Studia II stopnia

**Program:**

Informatyka

**Grupa przedmiotów:**

Przedmioty techniczne - zaawansowane

**Kod przedmiotu:**

MBI

**Semestr nominalny:**

3 / rok ak. 2021/2022

**Liczba punktów ECTS:**

4

**Liczba godzin pracy studenta związanych z osiągnięciem efektów uczenia się:**

1. liczba godzin kontaktowych – 47 godz.,w tym:
a. obecność na wykładach: 30 godz.,
b. udział w zajęciach komputerowych: 15 godz.,
c. udział w konsultacjach wykładowych oraz konsultacjach związanych z realizacją zajęć komputerowych: 2 godz.;
2. praca własna studenta – 72 godz., w tym:
a. przygotowanie do wykładów (przejrzenie materiałów z wykładu i dodatkowej literatury, próba rozwiązania ćwiczeń domowych sformułowanych na wykładzie, wykorzystywanie programów demonstracyjnych): 16 godz.,
b. przygotowanie do wykonania analiz, interpretacja wyników: 30 godz. (zapoznanie się z instrukcjami do ćwiczeń, instalacja oprogramowania, zapoznanie się z dokumentacją użytkownika, pobranie danych, analiza i interpretacja wyników, przygotowanie sprawozdania, omówienie wyników).
c. przygotowanie do egzaminu: 16 godz.
Łączny nakład pracy studenta wynosi: 120 godz., co odpowiada 4 pkt. ECTS.

**Liczba punktów ECTS na zajęciach wymagających bezpośredniego udziału nauczycieli akademickich:**

1,5 pkt. ECTS, co odpowiada 47 godz. kontaktowym

**Język prowadzenia zajęć:**

polski

**Liczba punktów ECTS, którą student uzyskuje w ramach zajęć o charakterze praktycznym:**

1,5 pkt. ECTS, co odpowiada 45 godz. realizacji zajęć komputerowych

**Formy zajęć i ich wymiar w semestrze:**

|  |  |
| --- | --- |
| Wykład:  | 30h |
| Ćwiczenia:  | 0h |
| Laboratorium:  | 0h |
| Projekt:  | 0h |
| Lekcje komputerowe:  | 15h |

**Wymagania wstępne:**

l umiejętność obsługi systemu Linux, w tym instalacji nowych pakietów;
l znajomość algorytmów i struktur danych, w tym algorytmów grafowych i algorytmów wyszukiwania napisów;
l umiejętność programowania w języku Python;
l znajomość podstaw probabilistyki i statystyki.

**Limit liczby studentów:**

60

**Cel przedmiotu:**

Celem przedmiotu jest zapoznanie słuchaczy z algorytmami stosowanych do analizy napisów reprezentujących sekwencje DNA, RNA i białka. Analizy takie są bardzo istotne w biologii i medycynie spersonalizowanej.

**Treści kształcenia:**

Wykład:
1. Wprowadzenie (2 godz.)
Bioinformatyka jako dziedzina informatyki. Rola analiz sekwencji we współczesnej biologii i medycynie. Budowa cząsteczek DNA, RNA i białek, reprezentacja tych cząsteczek jako napisów nad skończonym alfabetem, budowa genomu. Podstawowe reakcje inżynierii genetycznej.
2. Badanie podobieństw sekwencji biologicznych (4 godz.)
Programowanie dynamiczne, uliniowienie dwóch sekwencji, podobieństwo globalne i lokalne, algorytmy przybliżone, algorytmy o liniowym koszcie pamięciowym, algorytmy BLOSUM i PAM do obliczania macierzy podobieństwa symboli, algorytmy z afiniczną funkcją kary. Algorytmy do badania podobieństw wielu sekwencji. Profile. Wyszukiwanie motywów. Mediana napisów.
3. Bazy sekwencji biologicznych (2 godz.)
Wyszukiwanie sekwencji w bazie. Algorytmy heurystyczne FASTA, BLAST i pochodne. Formaty rekordów: FASTA, FASTQ. Istotność wyników. Podstawowe bazy sekwencji.
4. Asembling de-novo, re-sekwencjonowanie (6 godz.)
Sekwencjonowanie, sekwenatory 1, 2, i 3-ciej generacji. Kontig sekwencyjny i kontig fizyczny. Algorytmy oparte o graf pokrycia. Algorytmy oparte o pod-grafy grafów de Brujna. Algorytmy dla sprawowanych końców. Błędy odczytu. Sekwencje powtarzające się. Algorytmy stosowane do łączenia odczytów o różnej charakterystyce błędów. Algorytmy do tworzenia kontigów fizycznych. Miary jakości asemblerów DNA. Genom referencyjny. Mapa fizyczna i genetyczna. Sekwencje kodujące i niekodujące. Resekwencjonowanie.
5. Analiza genomu człowieka, analiza wariantów, choroby genetyczne (6 godz.)
Transformata Burrowsa-Wheelera, pliki SAM i BAM. Analiza wariantów genetycznych. Rodzaje chorób genetycznych. Rzadkie choroby genetyczne. Analiza wariantów. Plik VCF. Znajdowanie wariantów istotnych. Potoki w bioinformatyce. Analizy oparte o głębokość pokrycia. Wykrywanie zmian strukturalnych. Analizy oparte o markery genetyczne. Zmienność ludzkiego genomu. Markery STR i SNP. Badanie pokrewieństw. Badanie mieszanin DNA. Analiza haplotypów.
6. Drzewa filogenetyczne (2 godz.)
 Tworzenie drzew w oparciu o odległość sekwencji: metoda średnich połączeń, metoda przyłączania sąsiadów; tworzenie drzew w metodach bazujących na analizie symboli: metoda parsymonii, metoda największej wiarygodności.
7. Analizy oparte o ukryty model Markowa (2 godz.)
Łańcuchy Markowa. Ukryty model Markowa. Problem dekodowania. Algorytm Viterbiego. Algorytm prefiksowy i sufiksowy. Estymacja parametrów modelu Markowa. Algorytm Bauma-Welcha.
8. Analizy danych wielowymiarowych (2 godz.)
Grupowanie, Metody redukcji wymiarów, algorytm analizy składowych głównych.
9. Biologia syntetyczna i obliczenia realizowane na cząsteczkach DNA (4 godz.)
Struktura drugorzędowa biopolimeru, reprezentacja cząsteczki jako graf. Algorytmy obliczania struktury drugorzędowej na podstawie sekwencji: algorytm Nussinov, algorytm Zuckera. Optymalizacja sekwencji sztucznej cząsteczki DNA. Biologia syntetyczna. Obliczenia realizowane przez cząsteczki DNA. DNA komputer.

Projekt polega na wykonaniu szeregu analiz sekwencji biologicznych używając otwartego oprogramowania w zespołach 2 osobowych. Każde zadanie zajmuje kilka godzin pracy przy komputerze typu PC z systemem Linux, zakładając wcześniejsze pobranie danych i instalację odpowiednich narzędzi. Zadanie można wykonać samodzielnie albo z pomocą i asystą prowadzącego, na maszynie wirtualnej odpowiednio skonfigurowanej.
l Asembling de-novo DNA. Pobranie sekwencji z ogólnodostępnej bazy danych, generowanie odczytów zawierających błędy, uruchomienia assemblera de-novo, generowanie statystyk opisujących wyniki, analiza wyników.
l Adnotacja DNA. Pobranie zbioru kontigów (wyjście assemblera de-novo), adnotacja strukturalna - znajdowanie części kodujących i niekodujących, adnotacja funkcjonalna wykorzystując podobieństwo do opisanych elementów w bazach danych, analiza wyników.
l Resekwencjonowanie. Pobranie sekwencji chromosomu ludzkiego z ogólnie-dostępnej bazy danych, pobranie genomu referencyjnego, generowanie odczytów, mapowanie odczytów na genom referencyjny, znajdowanie wariantów.
l Analiza wariantów. Pobranie listy wariantów genetycznych oraz zbioru odczytów, analizy związane z głębokością pokrycia, wykrywanie zmian strukturalnych, szeregowanie znalezionych zmian uwzględniając ich istotność.

**Metody oceny:**

Realizacja przedmiotu obejmuje następujące formy zajęć:
l wykład prowadzony w wymiarze 2 godz. tygodniowo,
l zajęcia komputerowe realizowane samodzielnie w zespołach lub z asystą prowadzącego,
l konsultacje.
Aktywizacji studentów służą:
l interaktywna formuła wykładu,
l dostępność kilkudziesięciu algorytmów na stronie przedmiotu, które pozwalają wykonywać obliczenia krok po kroku wykorzystując przeglądarkę z językiem JavaScript dla własnych danych,
l dostępność terminu dla każdego z ćwiczeń, gdzie studenci mogą przyjść z własnym komputerem przenośnym i skonsultować wyniki z prowadzącym lub wykonać analizy pod opieką prowadzącego zajęcia komputerowe,
l wymóg przedstawienia sprawozdań z wykonanych ćwiczeń,
l omawiane narzędzia i dane są dostępne i bezpłatne.

Sprawdzanie założonych efektów kształcenia realizowane jest przez:
l ocenę wiedzy i umiejętności wykazanych na sprawdzianie pisemnym podczas sesji egzaminacyjnej,
l ocenę wiedzy i umiejętności związanych z realizacją analiz na zajęciach komputerowych.

**Egzamin:**

tak

**Literatura:**

1. Jin Xiong, Podstawy bioinformatyki, PWN, 2011.
2. R.Durbin, S.Eddy, A.Krogh, G.Mithison, Biological sequence analysis. Cambridge 2007.
3. P.Higgs, T.Attwood, Bioinformatyka i ewolucja molekularna, PWN, 2008.
4. Wing-Kin Sung, Algorithms for next-generation sequencing, CRC Press 2017.
5. V. Makinen, D. Belazzougui, F. Cunial, A. Tomescu, Genome-Scale Algorithm design, Cambridge 2015.
6. Pakiety języka Python.

**Witryna www przedmiotu:**

https://usosweb.usos.pw.edu.pl/kontroler.php?\_action=katalog2/przedmioty/pokazPrzedmiot&prz\_kod=103A-INxxx-MSP-MBI

**Uwagi:**

(-)

## Charakterystyki przedmiotowe