МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Львівський національний університет імені Івана Франка
Біологічний факультет
Кафедра генетики та біотехнології

Затверджено
на засіданні кафедри генетики та біотехнології
біологічного факультету
Львівського національного
університету імені Івана Франка
(протокол № 6 від 15 березня 2023 р.)

Завідувач кафедри._____
проф. Федоренко В.О

Силабус з навчальної дисципліни
"Біоінформатика"
що викладається в межах ОПП "Лабораторна діагностика біологічних систем"
другого (магістерського) рівня вищої освіти для здобувачів
за спеціальністю 091 Біологія та біохімія

Львів 2023

Силабус курсу Біоінформатика

Назва дисципліни	Біоінформатика
Статус дисципліни	Нормативна
Обсяг дисципліни	Лекції – 14 год, лабораторні – 14 год, самостійна робота – 122 год; разом – 150 год (5 кредитів ECTS)
Викладається для	Магістрів біології та біохімії, заочна форма, 2 (літній) і 3-й (осінній) семестри навчання
Форма контролю	Іспит
Галузь знань	Галузь знань – 9 Біологія, спеціальність – 091 Біологія та біохімія
Адреса викладання	Біологічний факультет, вул. Грушевського 4, Львів 79005
Розклад занять	https://bioweb.lnu.edu.ua/students/rozklad-ispytiv
Викладач (-і)	Богдан Омелянович Осташ (лекції) Юрій Васильович Ребець (лабораторні заняття)
Профайли викладачів	http://bioweb.lnu.edu.ua/employee/ostash-b-o https://www.researchgate.net/profile/Yuriy-Rebets
Контактний тел.	032 2394407
E-mail:	bohdan.ostash@lnu.edu.ua yurko.rebets@gmail.com
Сторінка курсу на сайті кафедри	https://bioweb.lnu.edu.ua/course/bioinformatyka
Консультації	*Очні консультації*: II семестр (2023 р), щовівторка, 11:30-13:00 *Онлайн- консультації:* у форматі "питання-відповідь" через електронну пошту, в робочі дні тижня, з 10:00-16:00; очікуйте на відповідь не пізніше ніж за три доби з моменту надходження питання

1. Коротка анотація до курсу

Станом на літо 2023 року в біомедичних базах даних є інформація про гени більше 400 000 організмів,або більш ніж 21 трильйон пар нуклеотидів (п.н.). Це колосальна кількість інформації про генетичну основу життя, і її кількість невпинно зростає. Нуклеотидні послідовності не мають змісту доти, доки дослідник його не визначить, як-от здатність кодувати білок, бути промотором тощо. Систематичний аналіз усього масиву даних експериментальними методами нереалістичний як економічно так і технологічно. Розширення таких баз даних стимулювало розвиток методів комп'ютерного аналізу нуклеотидних і амінокислотних послідовностей (НАП). Ці методи допомогли упорядкувати дані у межах різних спеціалізованих веб-сервісів і класифікувати їх відповідно до різних критеріїв і потреб дослідників. Однак найважливішим результатом синтезу молекулярної біології, геноміки, інформатики, комп'ютерних технологій і статистики – біоінформатики – стало те, що ця нова галузь перетворилася у самостійне знаряддя наукового відкриття. Курс "Біоінформатика" включає розділи, присвячені структурам баз даних; попарному і множинному вирівнюванню НАП; моделям, що узагальнюють множинні вирівнювання; основам філогенетичної реконструкції на основі; передбаченню функцій генів і пошукові мотивів у НАП; передбаченню тривимірних структур білків; аналізові молекул РНК.

Ключові слова: математичні моделі в біології, множинні і попарні вирівнювання, моделі Маркова, філогенетичні реконструкції, бази НАП.

2. Мета та цілі курсу

Мета: сформувати у слухачів курсу систему знань про основні бази даних нуклеотидних й амінокислотних послідовностей (НАП) та їхнього комп'ютерного аналізу, зокрема попарного і множинного вирівнювання, моделей на основі вирівнювань, філогенетичної реконструкції на основі НАП.

Цілі: *а*) викласти концептуальні математичні та молекулярно-біологічні засади, на яких ґрунтується біоінформатика; *б*) ознайомити студентів з наявним арсеналом методів біоінформатики, їхніх можливостей і обмежень, останніх тенденцій розвитку дисципліни; *в*) сформувати у студентів арсенал активних знань у галузі цієї дисципліни, тобто навчити їх формулювати наукові питання, відповіді на які можна шукати із залученням методів біоінформатики.

3. Формат курсу – очний або дистанційний

4. Результати навчання

Після курсу студент буде: *а*) знати основні бази даних НАП, принципи функціонування алгоритмів вирівнювання НАП, і спектр питань щодо структури, функції та еволюції НАП, на які можна шукати відповіді із застосуванням біоінформатичних веб-ресурсів і програм; *б*) мати базове розуміння методів філогенетичної реконструкції на основі НАП; *в*) вміти користуватись базами даних, де зберігаються нуклеотидні та амінокислотні послідовності, інформація про структури геномів, їхню експресію (транскриптомні дані – генні чіпи і RNAseq) та відповідні протеоми; *г*) вміти порівнювати НАП з гомологічними послідовностями за допомогою методів попарного і множинного вирівнювань; оперувати паттернами і профілями; *д*) вміти виявляти регуляторні послідовності у геномах – повтори, паліндроми, консервативні мотиви; передбачати функціональність гена (білка) на основі аналізу консервативних доменів і каталітичних центрів; *е*) вміти користуватись методами передбачення і моделювання структури білків; здійснювати молекулярно-філогенетичний аналіз.

Курс розроблено таким чином, щоб сформувати у студентів загальні і фахові компетентності:

ЗК01. Здатність працювати у міжнародному контексті.
ЗК02. Здатність використовувати інформаційні та комунікаційні технології.
ЗК03. Здатність генерувати нові ідеї (креативність).
ФК01. Здатність користуватися новітніми досягненнями біології, необхідними для професійної, дослідницької та/або інноваційної діяльності.
ФК02. Здатність формулювати задачі моделювання, створювати моделі об'єктів і процесів на прикладі різних рівнів організації живого із використанням математичних методів й інформаційних технологій.
ФК03. Здатність користуватися сучасними інформаційними технологіями та аналізувати інформацію в галузі біології і на межі предметних галузей.
ФК06. Здатність прогнозувати напрямки розвитку сучасної біології на основі загального аналізу розвитку науки і технологій.
ФК07. Здатність діагностувати стан біологічних систем за результатами дослідження організмів різних рівнів організації
ФК10. Здатність використовувати результати наукового пошуку в практичній діяльності.

Програмні результати навчання:

ПР1. Володіти державною та іноземною мовами на рівні, достатньому для спілкування з професійних питань та презентації результатів власних досліджень.

ПР2. Використовувати бібліотеки, інформаційні бази даних, інтернет ресурси для пошуку необхідної інформації.

ПР4. Розв'язувати складні задачі в галузі біології, генерувати та оцінювати ідеї.

ПР5. Аналізувати та оцінювати вплив досягнень біології на розвиток суспільства.

ПР6. Аналізувати біологічні явища та процеси на молекулярному, клітинному, організменному, популяційно-видовому та біосферному рівнях з точки зору фундаментальних загальнонаукових знань, а також за використання спеціальних сучасних методів досліджень.

ПР11. Проводити статистичну обробку, аналіз та узагальнення отриманих експериментальних даних із використанням програмних засобів та сучасних інформаційних технологій.

ПР14. Дотримуватись норм академічної доброчесності під час навчання та провадження наукової діяльності, знати основні правові норми щодо захисту інтелектуальної власності.

ПР16. Критично осмислювати теорії, принципи, методи з різних галузей біології для вирішення практичних задач і проблем.

5. Пререквізити та необхідне обладнання для вивчення курсу.

Знання англійської мови на рівні, достатньому для перекладу наукових статей; необхідні знання з основ генетики, біохімії, зоології та ботаніки. Розуміння базових математичних понять (логарифм, частка значень, експонента, відсоткові величини) та теорії імовірностей та статистичного аналізу даних. Базові навички роботи з комп'ютером. Наявність комп'ютера/смартфона з підключенням до інтернету (для лабораторних занять у випадку дистанційного формату навчання).

6. Політики курсу.

Відвідування лекційної частини курсу вільне. Матеріали лекційного курсу (PowerPoint-презентації) та електронний підручник буде надано електронною поштою усім студентам. Усі статті і матеріали, або гіперпосилання до них, що згадано нижче у схемі курсу (п. 7) – буде надано. Перша частина курсу (включно з лекцією про моделі на основі множинних вирівнювань – див. нижче) закінчується письмовим модулем. Написання модуля у визначений час обов'язкове, відсутність можлива лише за умови поважної причини, що має бути задокументовано (довідка про хворобу тощо). Відвідування лабораторних занять обов'язкове, під час яких студенти отримують бали за виконання контрольних завдань. Більше про систему оцінювання – див. нижче розділ **8**. Очікується, що студенти дотримуватимуться правил Академічної доброчесноті – див. http://www.lnu.edu.ua/wp-content/uploads/2019/06/reg_academic_virtue.pdf. Нульова толерантність (у вигляді недопуску до іспиту) до плагіату, списування, хабарництва. Зниження оцінки при виявленні фактів несамостійного підготовлення завдань до практичних занять (нерозуміння підготовленої презентації, механічне використання перекладів, згенерованих автоматичними перекладачами тексту).

7. Схема курсу

Лекція 1 (2 год). Порівняння НАП – концептуальні засади. Еволюційна спорідненість (гомологія) як концептуальна основа порівняння НАП. Гомологічність, подібність, ідентичність. Локальне і глобальне вирівнювання. Підпослідовності, прогалини, штрафи, рахунок вирівнювання. Еволюція НАП як процес Маркова. Матриці мутаційних даних PAM. Матриці BLOSUM. Емпіричні матриці кодонних заміщень і їхнє застосування в оцінці еволюції НАП. *Матеріали* – презентація лекції bioinf-L40.pdf. *Література*: [1, 2, 7].

Лабораторна 1 (2 год). Біоінформатичні сервіси на веб-порталі NCBI – PubMed, GenBank, Genome, Taxonomy, GEO datasets. Пошук інформації в PubMed. Ідентифікатори статей в PubMed – doi, PMID, PMCID. Підрозділ GenBank – архів нуклеотидних даних і продуктів їхньої трансляції. Структура файлів GenBank. Змінні і постійні ідентифікатори НАП в GenBank. Депонування нових НАП у GenBank. Підрозділ Genome – структура бази даних і її використання. Поняття геномного переглядача.

Самостійна робота (24 год). Ознайомлення з відкритим ресурсом для вивчення алгоритмів, що застосовуються в біоінформатиці - http://rosalind.info/problems/locations/. Основний фокус – на алгоритмах, що дають змогу аналізувати НАП – див. http://rosalind.info/problems/list-view/ . Що таке ДНК і білок. Центральна догма молекулярної біології XX століття, її сучасне тлумачення з точки зору епігенетики й теорії інформації. Біоінформатика як синтез методів молекулярної біології, генетики, інформатики і статистики. Маргарет О. Дейгоф і перші моделі еволюції НАП. Нуклеотид, кодон, амінокислотний залишок – елементарні одиниці інформації, якими оперує біоінформатика. Типи даних, що генерують геномні, транскриптомні і протеомні методи досліджень. Інтерактом. Системний аналіз. Роль біоінформатичних методів у біологічних дослідженнях. Журнал Nucleic Acids Research – провідник у світі біоінформатики. *Матеріали* – презентація лекції bioinf-L10.pdf. *Література*: [1, 2, 4] (див. список наприкінці схеми курсу).

Самостійна робота (10 год). Продовження вивчення можливостей пакета UGENE. Принцип графічної ілюстрації попарного вирівнювання НАП. Типи перебудов НАП, які можна виявляти за допомогою дотплот-аналізу – повтори, повні і часткові інверсії. Поняття про "вікно" вирівнювання. Приклади програм відкритого типу для дотплот-аналізу на рівні окремих генів і геномів.

Лекція 2 (2 год). Попарне вирівнювання НАП. Методи динамічного програмування у вирівнюванні НАП. Алгоритм локального вирівнювання Сміта-Уотермана з використанням унітарної матриці заміщень. Алгоритм глобального вирівнювання Нідельмана-Ванча. Порівняння рахунків вирівнювання НАП на основі унітарної матриці та BLOSUM62. *Матеріали* – презентація лекції bioinf-L50.pdf. *Література*: [1, 2, 3].

Лабораторна робота 2 (2 год). Попарне вирівнювання – наявні сервіси. Сторінка уведення даних для програми BLAST. Результати програми BLASTP.

Самостійна робота (10 год). Ознайомлення з пакетом програм попарного вирівнювання що базуються на алгоритмах динамічного програмування глобального і локального вирівнювання: https://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml . Як запит для білок-білкових вирівнювань використайте послідовність за № доступу GAW38200.

Лекція 3 (2 год). Веб-сервіс BLAST. Евристичні модифікації алгоритму локального попарного вирівнювання, що лежать в основі BLAST (Basic Local Alignment Search Tool) – "засівні слова порівняння", афінні штрафи, пороги подібності. Статистична оцінка результатів BLAST – *e*, *p*, *bits*, *gaps*. Родина програм BLAST – blastn, blastp, blastx, tblastn. PSI-BLAST – метод порівняння "профілів" білків. Структура початкової сторінки BLAST, її параметри за замовчуванням і можливості налаштування відповідно до мети дослідження. Структура сторінки результатів BLAST. Приклади вирівнювання високоподібних і віддалених НАП. *Матеріали* – презентація лекції bioinf–L60.pdf. *Література*: [1, 2].

Самостійна робота (10 год). Робота з алгоритмами blastx, blastn на веб-порталі NCBI

Лекція 4 (2 год). Множинне вирівнювання і основні моделі НАП на їхній основі. Концепція множинних вирівнювань НАП. Інформація, яку надає множинне вирівнювання НАП. Глобальні і локальні множинні вирівнювання. Веб-сервіси, що надають послугу множинного вирівнювання – CLUSTAL W2/Ω, MUSCLE, T-COFFEE. *Матеріали* – презентація лекції bioinf–L70.pdf. *Література*: [1, 2].

Лабораторна робота 3 (2 год). Алгоритм blastx. Статистика попарних вирівнювань. Налаштування сторінки пошуку для виявлення малоподібних послідовностей.

Самостійна робота 9 (15 год). Поглиблене ознайомлення з алгоритмічними основами множинних вирівнювань https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/download/lectures/PCB_Lect05_Multip_Align.pdf. Референтні бази множинних вирівнювань http://www.lbgi.fr/wscoperr?Balibase&FileMoi&macsimHtml&BB11033. Веб-портал HHSuite

Лекція 5 (2 год). Узагальнюючі моделі множинних вирівнювань – консенсусний рядок, паттерни. Синтаксис паттернів. PROSITE. Прості профілі, паттерни і позиційно-специфічні матриці (PSSM/PSWM). Поняття зваженого рахунку позиції вирівнювання і псевдорахунку. Бази PSSM – CDD. Генералізовані профілі. Концепція стану ознаки. Видимий шлях символів і прихований шлях станів. Приховані моделі Маркова (HMM) і сервіси на їхній основі. Pfam. *Матеріали* – презентація лекції bioinf-L80.pdf, bioinf-L90.pdf. *Література*: [1, 2, 3, 4].

Лабораторна 4 (2 год). Сервіси множинного вирівнювання – MUSCLE, T-COFFEE.

Самостійна робота (15 год). Ознайомлення з базою PROSITE, сервісом Conserved Domain Database https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml Ознайомлення з програмами попарного вирівнювання, що функціонують на основі HMM – HMMER3 (https://www.ebi.ac.uk/Tools/hmmer/), HHPred (https://toolkit.tuebingen.mpg.de/tools/hhpred). Як запит для білок-білкових вирівнювань використайте амінокислотну послідовність за номером доступу ADL32277.

Модульний контроль (2 год). Письмовий контроль за змістом перших 5 лекцій курсу.

Самостійна робота (10 год). Ознайомлення antiSMASH – HMM-опосередкованим сервісом виявлення генів вторинного метаболізму у геномах бактерій: https://antismash.secondarymetabolites.org/#!/start. Як запит для пошуку використайте номер доступу до нуклеотидної послідовності генома S. albidoflavus J1074 – NC_020990.

Лекція 6 (2 год). Реконструкція філогенії на основі НАП. Концепція філогенетичного дерева, її біологічний зміст. Основні терміни – клада, нода, корінь, аутгруп, шкала дивергенції. Філогенетичний сигнал. Матеріал для аналізу – нуклеотидні, кодонні чи амінокислотні послідовності? Стратегії вибору масиву даних для філогенетичного аналізу й тлумачення результатів. Гомологи, паралоги, ортологи. Еволюційна модель у філогенетиці. Дистанційні і позиційні методи філогенетичного аналізу. Метод "з'єднання сусідів" (NJ). Метод максимальної вірогідності (ML). Статистична оцінка достовірності отриманих філогенетичних дерев – метод бутстрап-аналізу для методу NJ і aLRT – для ML. Філогеномний аналіз і систематика життя. Значення філогенетичних підходів у популяційній генетиці і судовій практиці. Аналіз 16S рРНК. Філогенетичний веб-сервер Phylogeny.fr. *Матеріали* – презентація лекції bioinf-L100.pdf. *Література*: [1, 2, 6, 7].

Лабораторна 5 (4 год). Вступ до філогенетичного аналізу. Вибір даних, виявлення моделі еволюції. Філогенетична реконструкція на сервері phylogeny.fr. Дерево RuBisCo.

Самостійна робота (18 год). Як еволюція і філогенетика стосується мого повсякденного життя? – кожен зі студентів групи має вибрати і прочитати по одній статті зі списку, що є на веб-ресурсі університету Берклі: https://evolution.berkeley.edu/evolibrary/search/topics.php?topic_id=15.

Лекція 7 (2 год). Аналіз структур білків та РНК. Класифікація білків. Поняття родини і фолду. Бази даних Pfam, SCOP. Тривимірні моделі білків – яку інформацію вони містять? PDB. Програма пошуку структурної гомології – HHPred. Веб-сервер ExPaSy для визначення основних параметрів білкових послідовностей та імовірних ділянок їхнього протеазного розщеплення і посттрансляційної модифікації. Програми для моделювання третинної структури білків і докінгу малих молекул. Веб-сервер STRING для аналізу функції гена у всій сукупності зв'язків з сусідніми генами і спорідненими геномами. KEGG. AlphaFold. Виявлення рРНК й тРНК у геномах. Аналіз даних RNAseq. Бази даних тРНК. Передбачення вторинної структури РНК та оцінка її стабільності. Бази даних рРНК для потреб молекулярної таксономії. Бази даних некодуючих РНК. Бази даних виявлення CRISPR-елементів у геномах бактерій. *Матеріали* – презентація лекції bioinf-L120.pdf, bioinf-L130.pdf. *Література*: [1, 2, 6, 7].

Лабораторна 6 (2 год). Бази тривимірних структур білків. PDB. Phyre2 (http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index) – сервер передбачення тривимірних структур білків. Сервіси передбачення вторинних структур РНК - http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi

Самостійна робота (10 год). Ознайомлення з життям і науковим доробком українсько-американського вченого-генетика Теодозія Добжанського – https://www.youtube.com/watch?v=TH2AC8fu34M. Сумісність біблійного та еволюційного вчень – погляд Т. Добжанського https://www.pbs.org/wgbh/evolution/library/10/2/text_pop/l_102_01.html

Література

1. Осташ Б.О. Біоінформатика: аналіз генетичних послідовностей. Електронний підручник. Видавничий центр ЛНУ ім. Івана Франка, 2022, 232 стор. ISBN 978-617-10-0729-1. Доступ онлайн: http://dspace.lnulibrary.lviv.ua/handle/123456789/169

2. Higgs PG, Attwood TK. Bionformatics and Molecular Evolution. Blackwell Publishing, Oxford, 2005. 398 p. ISBN 1–4051–0683–2.
3. Durbin R, Eddy S, Krogh A, Mitchison G. Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, 1998. 371 p. ISBN-13 978-0-521-62971-3
4. Borodovsky M, Ekisheva S. Problems and Solutions in Biological Sequence Analysis. Cambridge University Press, Cambridge, 2006. 362 p. ISBN-13 978-0-521-61230-2
5. Bioinformatics: a practical guide to the analysis of genes and proteins, 2nd Ed / AD Baxevanis, BFF Ouellette. – New York: John Wiley & Sons, 2001. – 455 p.
6. Allman ES, Rhodes JA. Mathematical Models in Biology. An Introduction. Cambridge University Press, Cambridge, 2003. 386 p.
7. Pevsner J. Bioinformatics and functional genomics. 3rd edition. Wiley Blackwell, London. – 2015- 1116 p. ISBN 978-1-118-58178-0.

8. Система оцінювання та вимоги

Загальна система оцінювання курсу	участь в роботі впродовж семестру/іспит - 50/50
Вимоги до письмової роботи (модуль)	За змістом перших дев'яти лекцій буде виконано поточний контроль знань у вигляді написання модуля. В модуль входять: визначення термінів (10 балів), два питання (по 6 б. кожне), одна схема чи таблиця, яку треба заповнити/зобразити (8 б.). Максимальна оцінка за модуль – 20 балів. Написання модуля обов'язкове.
Лабораторні заняття	Ще 30 балів студент може набрати упродовж семестру за виконання шести контрольних завдань упродовж лабораторних занять (Бази даних, BLAST, MSA, WebLogo, Phylo1, Phylo2).
Умови допуску до підсумкового контролю	Сумарний бал за модуль і за виконання контрольних завдань під час лабораторних занять має становити не менше 25, у такому співвідношенні: не менше 5 за модуль і не менше 20 за лабораторні.
Іспит	50 балів. Набір питань аналогічно до модуля; до термінів, питань і схем додаються тести. Письмова підготовка на протязі не більше 30 хв, далі усна відповідь. На іспит виноситься весь матеріал курсу

9. Навчальні методи

- Словесні (лекції, дискусії, пояснення)

- Практичні (лабораторні заняття)

- За типом пізнавальної діяльності – проблемно пошукові, репродуктивні.

10. Перелік питань і типових задач, що виноситься на іспит

1. Моделі, що описують частоту зустрічності "слів" у генетичних послідовностях – Бернуллі і ланцюги Маркова.
2. PubMed – стратегії пошуку бібліографічних даних, значення ідентифікаторів статей
3. Структура флет-файлу GenBank
4. Національний центр біотехнологічної інформації США (NCBI) – структура і функції
5. Теорія прийнятних точкових мутацій (PAM) М. Дейгоф
6. Опишіть процес побудови позиційно-специфічних рахункових матриць (PSSM).
7. Алгоритми динамічного програмування у попарному вирівнюванні.
8. Основні терміни і теоретичні засади попарного вирівнювання
9. Опишіть етапи прогресивного принципу множинного вирівнювання
10. Опишіть основні елементи сторінки результатів BLAST
11. Підходи до уведення прогалин у попарні й множинні вирівнювання
12. У рівнянні з теореми Карліна-Альтшуля число очікування E прямо пропорційне простору пошуку (розміру бази даних) і обернено рахунку вирівнювання S. Тобто, у більших базах імовірність виявити випадкові збіги мала б зростати. Однак на практиці більші бази даних приводять до вирівнювань з дуже низьким E. Чому?
13. Які підходи використовують до оцінювання розривів (прогалин) у попарних вирівнюваннях?
14. Основні параметри опису попарного вирівнювання: S, S', E, P, ідентичність ID, подібність SI
15. Параметри налаштування програми BLAST
16. PAM-матриці і спосіб їхньої побудови
17. Опишіть принципи побудови механістичних та емпіричних матриць заміщень; що відмінного і спільного між ними.
18. Принцип побудови матриць BLOSUM, їхні відмітні властивості (порівняно з PAM).
19. Відмінні і спільні риси паттернів, позиційно-специфічних матриць і прихованих моделей Маркова
20. Чому не існує не одна, а серія матриць PAM або BLOSUM? Як обчислюють матрицю PAM250? Для яких потреб слід використовувати матрицю PAM20?
21. Основні властивості моделей еволюції нуклеотидних послідовностей.
22. Який біологічний зміст несе уведення розривів (прогалин) у попарні вирівнювання? Чи можна переставляти місцями позиції вирівнювання, і якщо так – то який біологічний процес відображає така маніпуляція НАП
23. Принципи дотплот-аналізу НАП
24. Що спонукало дослідників до розробки емпіричних підходів до оцінки вирівнювань амінокислотних послідовностей? Чому емпіричні підходи не набули поширення для нуклеотидних послідовностей?
25. Спільне і відмінне в попарному і множинному вирівнюваннях
26. Стисло опишіть моделі, які створюють на основі множинного вирівнювання НАП. Що в них відмінного?
27. Еволюційні засади попарного вирівнювання. Основні терміни.
28. Поясніть, що таке позиційно-незалежні та позиційно-специфічні рахунки вирівнювання, наведіть приклади їхнього використання в аналізі НАП
29. Концепція псевдорахунків у біоінформатиці – приклади її використання
30. Опишіть відомі вам способи множинного вирівнювання НАП

31. Принцип функціонування алгоритму пошуку оптимального глобального вирівнювання (Нідельмана-Ванча)
32. Відмінність між алгоритмами глобального і локального вирівнювання НАП
33. За якими ознаками певну модель чи процес можна віднести до Марковського?
34. Яку роль відіграють константи λ й K в рівнянні для обчислення числа E?
35. Етапи побудови позиційно-специфічної рахункової матриці
36. Моделі оцінки частот символів/слів у НАП
37. Вичерпні та евристичні підходи до попарного вирівнювання НАП.
38. Порівняйте позиційно-специфічні матриці й приховані моделі Маркова як методи опису множинних вирівнювань.
39. ДНК- і білкові логотипи – принцип побудови, та інформація, яку він містить
40. У чому полягає складність побудови множинних вирівнювань? Які є способи оцінки їхньої якості?
41. Принцип функціонування алгоритму BLAST.
42. Опишіть, як зміна налаштувань програми BLAST впливає на результат пошуку гомологів?
43. Прогресивний та ітеративний принципи множинного вирівнювання
44. Основні етапи побудови і використання позиційно-специфічних вагових матриць
45. Можливості й обмеження методів попарного і множинного вирівнювання для виявлення гомологічних НАП
46. Які дані містить сторінка результатів BLASTP?
47. Основні етапи пошуку оптимального локального вирівнювання за алгоритмом Сміта-Уотермана
48. Основні поняття у галузі філогенетичної реконструкції
49. Основні елементи філогенетичного дерева, взаємозв'язок між ними
50. Вибір даних для філогенетичної реконструкції
51. Що таке модель еволюції у філогенетичній реконструкції і яке її значення?
52. Опишіть основні етапи філогенетичної реконструкції
53. Методи статистичної оцінки філогенетичних дерев
54. Підхід до філогенетичної реконструкції: з'єднання сусідів (NJ)
55. Підхід до філогенетичної реконструкції: максимальної ощадності (MP)
56. Підхід до філогенетичної реконструкції: максимальної вірогідності (ML)
57. Чому дерево-провідник з прогресивних підходів до множинного вирівнювання не є філогенетичним деревом?
58. У чому полягає суть курування вихідного множинного вирівнювання, що передує філогенетичній реконструкції?
59. На чому ґрунтується суть пошуку мотивів ДНК?
60. Поясніть поняття родина, фолд, клас білка (за системою SCOP)
61. Поясніть принципи пошуку гомологів за допомогою програм BLAST й HHPred
62. Чому порівняння первинних і третинних структур генетичних послідовностей дає різний результат?
63. Практичне використання філогенетичного аналізу

Задачі

1. Прочитайте такий паттерн: [HTS]-C-x-{P}-C-x(2)-C-{CP}-x(2)-C-[PEG]

Що можна сказати про частоти вживання амінокислотних залишків у першій, другій та останній позиціях вирівнювання, з якого побудовано паттерн?

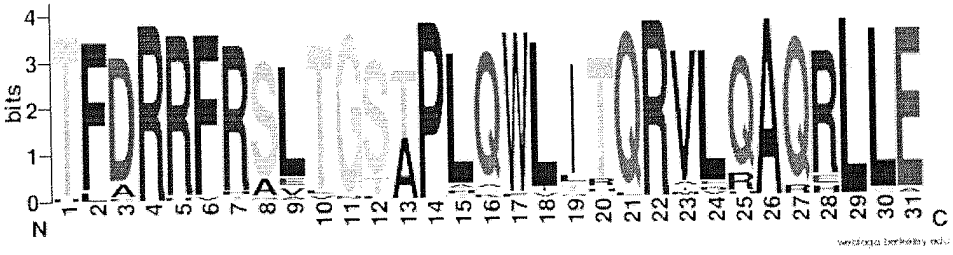
2. Рахунок 4 для амінокислотних залишків валіну (V) та ізолейцину (I) у матриці PAM250 відображає співвідношення цільової частоти M_{ij} (імовірність зустріти V й I

в одній позиції вирівнювання гомологічних білків) до фонової частоти V f_i (= імовірність, що ці два залишки будуть в одній позиції вирівнювання випадкових білків). Скориставшись рівнянням М. Дейгоф для обчислення S (№ 25 у лекціях і № 19 у конспекті), обрахуйте, чому дорівнює це співвідношення для випадку V-I. Напишіть, у скільки разів частіше пара V-I зустрічається у вирівнюваннях гомологічних білків ніж випадкових.

3. На еволюційній відстані 1PAM цільова частота заміщення аспартату (D) на залишок глутамату (E),M_{DE}, становить 0.0056, а на відстані 250PAM – 0.11. Обчисліть рахунок заміщення S_{DE} на відстанях 1PAM і 250PAM (фонова частота аспартату – 0.047). Виходячи з отриманих даних, вкажіть, на якій еволюційній відстані заміщення аспартату на глутамат можна вважати прийнятною мутацією.
4. На рисунку наведено заповнену матрицю для алгоритма Нідельмана–Ванча. Відстежте оптимальний шлях вирівнювання, зобразіть відповідне попарне вирівнювання і визначте його рахунок S.

		G	T	A	C	G	T	C	G	G
	0	3	-6	-9	-12	-15	-18	-21	-24	-27
A	3	5	-8	-2	-1	-4	-7	-10	-13	-16
T	-6	-8	3	0	-3	-6	-4	-1	-2	-5
A	-9	-11	0	11	8	5	-2	-1	-4	-7
C	-12	-14	-3	8	19	16	13	10	7	-4
A	-15	-17	-6	5	16	14	11	8	5	-2
T	-18	-20	-9	-2	13	11	22	19	16	13
G	-21	-16	-12	-1	10	21	19	17	27	24
T	-24	-13	-2	-4	-7	-18	-29	-26	-24	-22
C	-27	-16	-5	-7	-4	-15	-26	-37	-34	-31
T	-30	-19	-8	-10	-1	-12	-23	-34	-32	-29

5. Задано дві амінокислотні послідовності, ABCDEFG та ABCDEDEFG. Побудуйте дотплот-графік цих послідовностей. Яку генетичну перебудову ілюструє отриманий графік?
6. Задано дві амінокислотні послідовності: ALITTLE й ALITLE. Виконайте їхнє локальне попарне вирівнювання за двох режимів обчислення рахунків. Перший: збіг +1, незбіг 0, нема штрафів за розрив (прогалина = 0). Другий: збіг +1, незбіг 0, штраф за відкриття прогалини -3, за продовження 0. (Тут прогалина на кінці послідовності не рахується як розрив, а є незбігом – тому це локальне вирівнювання). Запишіть рахунки отриманих вами вирівнювань.
7. Розгляньте логотип НАП. Що він підсумовує? Які позиції абсолютно консервативні? Які містять приблизно однакові кількості двох різних амінокислот? Що означає вісь ординат?



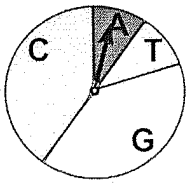
8. Внизу наведено попарне вирівнювання. Обчисліть його рахунок S за такою схемою: збіги +4, незбіги -3, відкриття прогалини -5, продовження -0.1. Чи можна ці послідовності вирівняти краще (отримати вищий рахунок)?

```
AGCTTCGAC-C
ACCTTCGACAC
```

9. Який мав би бути мінімальний біт-рахунок S' для пари послідовностей завдовжки 156 та 182 ак залишки, аби гарантувати відсутність випадкових вирівнювань з рахунком рівним або більшим S'?
10. Порівнюють дві амінокислотні послідовності завдовжки 200 ак. Яке очікуване число випадкових вирівнювань послідовностей такого ж розміру можна отримати з біт-рахунком $S' = 15$?
11. Гомолог заданої послідовності Y міститься у двох базах даних – SwissProt та PDB. Після виконання BLAST-пошуку у базі SwissProt вирівнювання Y з гомологом отримало значення $E = 1.5$. Таке саме вирівнювання при BLAST-аналізі бази PDB мало значення $E = 0.075$. Яке співвідношення розмірів вищезгаданих баз даних? Котра більша?
12. Порівнюють дві амінокислотні послідовності завдовжки 150 ак. Яким має бути біт-рахунок S' вирівнювання цих двох послідовностей, аби очікуване число випадкових вирівнювань послідовностей такого ж розміру з рахунком $\geq S'$ було 0.001 ?
13. Задану послідовність Б використано для BLAST-пошуку гомологів у базі GenBank. Отримано хіт, що вирівнюється з Б, і це вирівнювання описується такими значеннями: біт-рахунок $S' = 160$ число очікування $E = 1.88e\text{-}35$. Який розмір мав простір пошуку? Прийнявши, що розмір послідовності Б становить 300 ак залишків, визначте розмір бази GenBank.
14. Задану послідовність Б розміром 400 ак залишків використано для BLAST-пошуку гомологів у геномах ссавців. Отримано три хіти, з такими рахунками вирівнювання S й асоційованими з ними значеннями E: $S = 52$, $E = 0.011$; $S = 48$, $E = 0.055$; $S = 52$, $E = 0.011$; $S = 40$, $E = 1.352$. Які з вирівнювань можна вважати невипадковими? Відповідь обґрунтуйте. Обчисліть розмір бази даних, проти якої порівнювали послідовність Б, враховуючи що $K = 0.15$, $\lambda = 0.4$.
15. Яким має бути біт-рахунок S' вирівнювання заданої послідовності завдовжки 200 ак залишків із знайденою, виявленої у базі GenBank (96×10^9 залишків), аби це вирівнювання вважалося невипадковим?
16. Послідовність W використано для BLAST-пошуку гомологів у базі GenBank (98×10^9 ак залишків). Отримано хіт, що вирівнюється з W, з біт-рахунком S' 140 і числом очікування 2.8e-29. Який розмір послідовності W ?
17. Якщо попарне вирівнювання, отримане при аналізі бази даних, має число очікування 0.0, то його слід розцінювати як випадкове чи невипадкове? Відповідь обґрунтуйте.
18. Задану послідовність R 200 ак залишків завдовжки використано для BLAST-пошуку гомологів у базі розміром (80×10^9 ак залишків). Отримано хіт, що вирівнюється з R, і це вирівнювання має число очікування $E = 0.0035$. Який біт-рахунок S' має це вирівнювання?
19. Рисунок внизу зображує ланцюг Маркова, що генерує нуклеотидну послідовність із певним частотним розподілом символів (= нуклеотидів). Нехай задано дві послідовності:

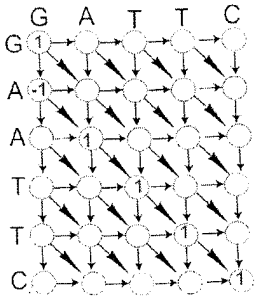
1: GATCGAATTCATTAATCTTA
2: GGGCACCTCCATGTTGGCCT

Яку з цих послідовностей з більшою вірогідністю генеруватиме наведена модель? Обґрунтуйте відповідь, визначивши імовірності генерування послідовностей 1 і 2 за умови цієї моделі



p_A=0.1, p_C=0.41, p_G=0.39, p_T=0.1

20. Рисунок внизу показує оптимальний шлях вирівнювання двох нуклеотидних послідовностей. Зобразіть це вирівнювання у текстовому (попарному) форматі. Обчисліть рахунок цього попарного вирівнювання (збіг: +1, незбіг – 0, прогалина -1)



21. На рисунку внизу зображено два попарні вирівнювання. Котре з них глобальне, а котре – локальне? Обчисліть рахунок вирівнювання обидвох за такою схемою: збіг -+2, незбіг 0, штраф за відкриття прогалини -1, за продовження -0.1 (в локальному вирівнюванні визначте рахунок лише ділянок, що перекриваються: t-c --- g-c).

```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC
  |  || |  ||  | | | |||     || | | |  | |||| |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG-T-CAGAT--C
```

```
                tccCAGTTATGTCAGggacacgagcatgcagagac
                   ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

22. Розгляньте два білкові логотипи. Які позиції абсолютно консервативні? Чому у першій позиції залишок валіну вищий (в обидвох логотипах) за залишок ізолейцину? Яка причина різної висоти літер у двох логотипах? Що означає вісь ординат?

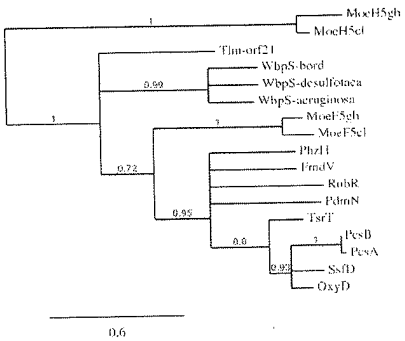
23. Послідовність S_0 складається із 100 п н. S_1, що походить з S_0, містить 11 транзицій і 9 трансверсій порівняно з предковою послідовністю. Обчисліть еволюційну відстань d між S_1 й S_0 за моделлю Джакса-Кентора (JC69).

24. Внизу ліворуч наведено попарне вирівнювання білкових послідовностей, яке виконано на основі такої системи обчислення рахунку вирівнювання S: збіги і незбіги – BLOSUM62, штраф за відкриття прогалини -5. Обчисліть S цього вирівнювання. За яких умов було б вигіднішим вирівнювання цих двох послідовностей у спосіб, що зображено внизу праворуч?

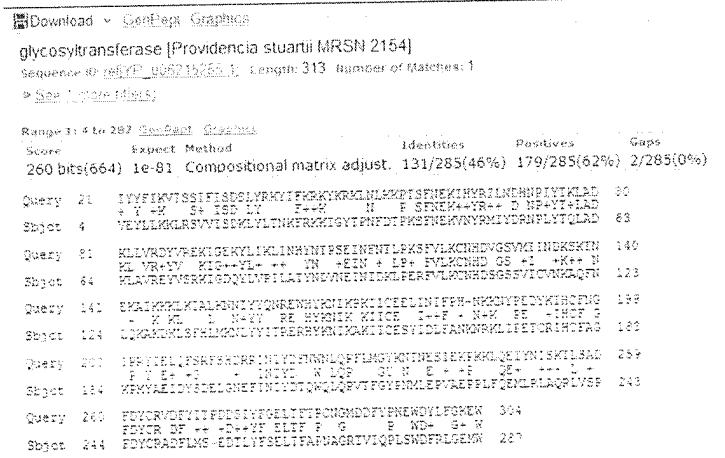
```
ALDGWTSP
ALEGPTSP
```

```
ALDGW-TSP
ALEG-PTSP
```

25. Дайте визначення основним елементам філогенетичного дерева, зображеного на малюнку



26. Внизу наведено типовий результат програми BLASTP. Поясніть зміст усіх термінів і позначень, наведених на ньому



Download ~ GenPept Graphics

glycosyltransferase [Providencia stuartii MRSN 2154]

Sequence ID: [illegible] Length: 313 Number of Matches: 1

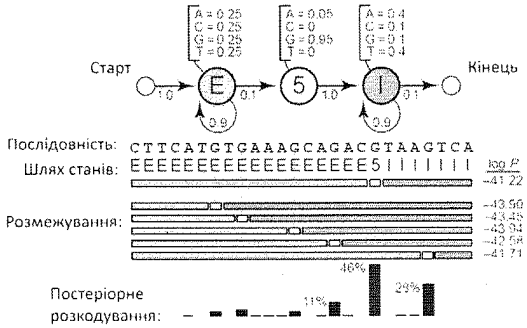
» See [illegible]

Range 1: 4 to 287 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
260 bits(664)	1e-81	Compositional matrix adjust.	131/285(46%)	179/285(62%)	2/285(0%)

```
Query  21   IYYFIKVTSSIFISDSLYRKYIFKRKYKRKLNLKKPTSFNEKIHYRILNDNNPIYIKLAD  80
            + Y +K   S+ ISD LY     F++K     N   P SFNEK++YR++ D NP+YI+LAD
Sbjct  4    YEYLLKKLRSYVISDKLYLTNKFRKKIGYTPNFDTPKSFNEKVNYRMIYDRNPLYIQLAD  63

Query  81   KLLVRDYVREKIGEKYLIKLINHINIPSEINFNTLPKSFVLKCNHDVGSVMIINDKSKIN  140
            KL VR+YV  KIG++YL+ ++  YN  +EIN + LP+ FVLKCNHD GS +I  +K++ N
Sbjct  64   KLAVREYVERKIGDQYLVFILATYNDVNEINIDKLPERFVLKCNHDSGSSVICVNKQQFN  123

Query  141  EMATKKKLMIALKNNIYYQNRENHYKNIKPKIICEELINIFPN-NKSNYPEDYKIRCFNG  199
            -  K KL   L  N+YY  RE HYKNIK KIICE  I++P - N+N  PE  -IRCF G
Sbjct  124  LDKAAKKLSFHLMKNLIYYITERRHYKNIKAKIICESYIDLFANKNRKLIPETCRIRCFAG  183

Query  200  IPRYIELIFSAFSKCRFINIYDFNKNLQPFLMGYKKTNESIEKFKKLQEIYNISKTLSAD  259
            P Y E+ +I    +  INIYI  N LQP    GY K  E + +E    QF+  +++ I +
Sbjct  184  KPKYAEIIVSDELGNEFINIYDTQKILQPVTFGYRKMLEFVAEFFLFIENLRLAQRLVSF  243

Query  260  FDYCRVDFYITPDESIYFGELITFPCNGMDDFVPNEWDVLFGKEW  304
            FDYCR DF ++ +D++YF ELIF P  G    E  ND+  G+ K
Sbjct  244  FDYCRADFLMS-EDTLIYFSELITFAPWAGRTVIQPLSWDFRLGEMW  287
```

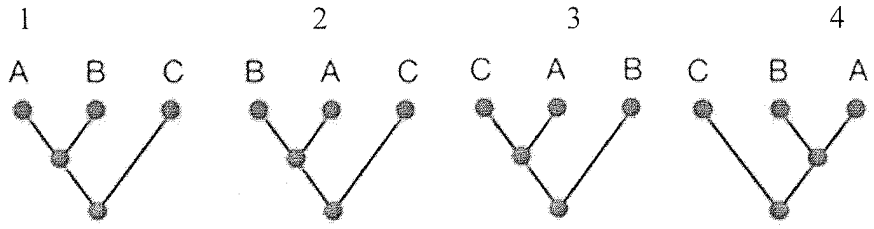
27. Поясніть, що зображено на малюнку, який наведено внизу. Як використовується математична модель, яку відображено на цьому малюнку?



28. Знайдіть оптимальний шлях вирівнювання за методом Нідельмана-Ванча для таких двох послідовностей

	C	O	E	L	A	C	A	N	T	H
P										
E										
L										
I										
C										
A										
N										

29. Внизу наведено 4 філогенетичні дерева, 1-4. Виберіть з них одне, що має топологію відмінну від решти трьох. Вибір поясніть



11. Опитування

Буде виконано в кінці курсу

Робочу програму навчальної дисципліни (силабус):
Складено д-ром біол. наук, доцентом Осташем Б.О.

Погоджено

Голова методичної ради біологічного факультету

_____ Віталій ГОНЧАРЕНКО

15.03. 2023 р.

Гарант ОПП

_____ Олена СТАСИК

14.03 2023 р