Міністерство освіти і науки України

Національний технічний університет України

«Київський політехнічний інститут»

С.В. Горобець, О.Ю. Горобець, Т.А. Хоменко

ОСНОВИ БІОІНФОРМАТИКИ

Підручник для студентів

напряму підготовки 6.051401 «Промислова біотехнологія»

факультету біотехнології і біотехніки

Рекомендовано Методичною радою НТУУ «КПІ»

(протокол № 4 від 23 грудня 2010 р)

Київ

НТУУ «КПІ»

2010

ВСТУП

З 50-х років минулого сторіччя в області молекулярної біології інтенсивно проводяться дослідження, спрямовані на отримання нуклеотидних послідовностей різних організмів, а також людини. Необхідність обробки отриманих експериментальних даних призвела до появи нового напрямку молекулярної біології – біоінформатики. Це складна дисципліна, що потребує від фахівців знань в області фізики, хімії, математики, біології та інформатики. Розвиток цієї галузі науки активно стимулюється та фінансуються фармацевтичними і біотехнологічними фірмами, а також науковими фондами і урядами США, Японії та країн Євросоюзу.

Біоінформатику відносять до числа високих технологій сучасної біології, що забезпечує інформаційно-комп'ютерні та теоретичні основи генетики і селекції, молекулярної біології, генетичної та білкової інженерії, біотехнології, медичної генетики, генної діагностики та екології. З точки зору біоінформатики молекули нуклеїнових кислот та білків представляють собою текст, що складається з обмеженої кількості алфавітних знаків (4 для нуклеїнових кислот та 20 для білків). Задачею біоінформатики є аналіз змісту цього тексту. Можна виділити три основних напрямки біоінформатики:

– створення баз даних, які дозволяють зберігати велику кількість класифікованих біологічних даних і керувати ними;

– розробка алгоритмів і методів статистичного аналізу для визначення співвідношень між тими чи іншими даними;

– використання програмного забезпечення, алгоритмів і баз даних для аналізу біологічних даних різних типів, зокрема, послідовностей ДНК, РНК і білків, білкових структур, профілів експресії генів і біохімічних шляхів, та отримання на основі цих даних нових знань.

Крім того, варто відзначити, що просторова структура макромолекул ДНК, РНК і білків є визначальною у формуванні механізмів їх взаємодії з іншими атомами та молекулами. Визначити тривимірну структуру біологічних макромолекул за допомогою електронної мікроскопії, рентгеноструктурного аналізу або методу ядерного магнітного резонансу складно, а часто і взагалі неможливо. Така ситуація призвела до інтенсивного пошуку біоінформаційних методів передбачення структури молекул. Тому визначення точної просторової структури біомолекул є однією з найголовніших задач біоінформатики.

Результати досліджень в галузі біоінформатики знаходять застосування при створенні нових лікарських препаратів та при пошуку хімічних речовин для стимуляції сільськогосподарського виробництва в залежності від генотипу певної рослини або тварини.

Даний підручник з основ біоінформатики допоможе студенту засвоїти основні задачі та методи цієї науки. Оскільки знання з біоінформатики мають міждисциплінарний характер, вони дозволять розв’язувати широке коло задач у області біотехнології, молекулярної біології, генетики та інформатики.

РОЗДІЛ 1

ПРЕДМЕТ, ЦІЛІ ТА ЗАДАЧІ БІОІНФОРМАТИКИ

1.1 Фактори, які сприяли розвитку біоінформатики.

За останні 25-30 років науковцями світу було накопичено колосальний експериментальний матеріал про будову і функціонування біологічних молекул (білків і нуклеїнових кислот). Цей матеріал потребує розвинутих комп'ютерних методів для свого аналізу, саме тому біоінформатика широко застосовується для розв'язання задач генетики, молекулярної біології, молекулярної біотехнології, біохімії та ін.

На сьогодні ці дисципліни є лідерами як за обсягом надходжень нової інформації, так і за темпами її впровадження в різні технології. Математикам, наприклад, від обчислення площі кола до створення перших обчислювальних машин знадобилося більше 2000 років, а фізикам від відкриття закону всесвітнього тяжіння до польотів у космос – біля 300 років. Молекулярній же біології від відкриття структури ДНК та принципів кодування генетичної інформації у 1953 році до її широкомасштабного промислового використання знадобилося всього біля 30 років.

У 1951 році – американський хімік і фізик К. Л. Полінг разом із американським біохіміком Р. Б. Корі розробили уявлення про структуру поліпептидного ланцюга у білках, вперше висунули гіпотезу про її спіральну будову та описали альфа-спіраль.

Подвійна спіраль ДНК була відкрита у 1953 році 25-річним американцем Джеймсом Уотсоном та 36-річним англійцем Френсісом Кріком, що поклало початок бурхливому розвитку генетики та молекулярної біології, який триває і в наші дні. У 1962 році відкривачі вторинної структури ДНК стали Нобелівськими лауреатами.

Першим розшифрованим білком був бичачий інсулін (1956 р.), що складався із 51 амінокислотного залишку. Приблизно десятиліттям пізніше

була розшифрована перша нуклеїнова послідовність тРНК дріжджів із 77 основ. У 1965 році Маргарет Дейхофф з співробітниками Національного Фонду біомедичних досліджень (Вашингтон) систематизувала усі наявні дані про амінокислотні послідовності і створила першу біоінформаційну базу даних – атлас білкових послідовностей та їх структур. Перша версія атласу містила опис 65 послідовностей. Через декілька років після розшифровки перших амінокислотних послідовностей були створені перші біоінформаційні (у той час – біологічні) бази даних білків і нуклеотидних послідовностей. А у 1982 році були організовані банки даних нуклеотидних послідовностей – GenBank (база послідовностей ДНК) у США та EMBL (Європейська молекулярно-біологічна бібліотека) у Європі.

Істотний прорив в області молекулярної біології пов’язують з відкриттям у 70-х роках 20 сторіччя зворотньої (оберненої) транскрипції (синтезу ДНК з використанням РНК в якості матриці) та клонування ДНК і створенням на цій основі методів швидкого секвенування, тобто із встановленням первинної структури нуклеотидних послідовностей. У 1987 р. було розроблено методи автоматичного секвенування (Л. Худ, Т. Хункапиллер). Водночас потік інформації стосовно структури та властивостей генів і кодованих ними білків почав зростати експоненційно (рис.1). Обробка та аналіз цієї інформації традиційними статистичними методами стали практично неможливими.

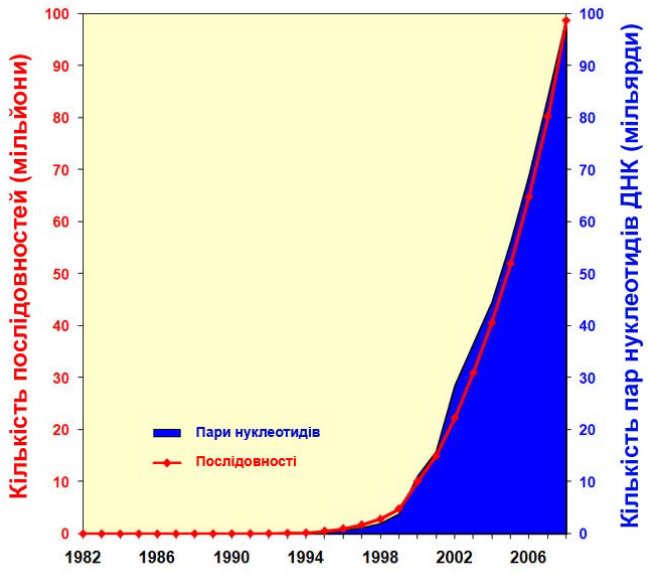


Рис. 1.1. Темпи зростання об’єму інформації в БД GenBank на протязі 1982-2008 років (станом на 03.02.2009, http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html).

Кожна розшифрована нуклеотидна або амінокислотна послідовність сама по собі представляє значний інтерес для генної інженерії та біотехнології, але наряду з цим, послідовність може слугувати колосальним джерелом інформації при порівнянні її з іншими послідовностями. Це стимулювало розробку добре структурованих баз даних (БД), які могли б працювати з великими об’ємами експериментальних даних, і спеціальних програмних засобів, що допомогли б інтерпретувати результати експериментів. Таким чином, у 80-ті роки минулого сторіччя виникла нова наука – біоінформатика.

Відомості про відкриття нових послідовностей розміщувалися у провідних журналах, а потім ці дані заносилися до БД вручну. Однак, коли почалося лавиноподібне зростання обсягів інформації, такий процес став неможливим. Журнали почали вимагати, щоб послідовності розміщувалися у БД самими авторами. Сьогодні, коли секвенування ДНК є достатньо поширеним процесом, який можуть виконувати роботи або студенти на лабораторних роботах, багато послідовностей можуть потрапляти до БД без опублікування у наукових журналах. На даний час існують такі БД, де інформацію може розмістити кожен науковець, і такі, де інформація суворо перевіряється, а відповідальність за її достовірність покладається на власника бази даних. БД постійно

обмінюються інформацією, але засоби роботи з даними, що розробляються, наприклад, у Центрі біотехнологічної інформації США і Європейському інституті біоінформатики відрізняються між собою.

Вважається, що першим важливим з біологічної точки зору результатом, який було отримано за допомогою аналізу послідовностей, було виявлення подібності вірусного онкогену v-sis і нормального гену фактора росту тромбоцитів, що сприяло значному прогресу у розумінні механізму раку. Відтоді робота з послідовностями стала необхідним елементом біологічних досліджень.

На початку нового тисячоліття відбулась визначна подія в науковому та суспільному житті – була прочитана повна нуклеотидна послідовність генома людини, що сприяло значному розвитку геноміки. Геноміка - наука, яка займається інтегральним дослідженням генів, аналізом їх структури і функції, особливостями і послідовностями регуляторних ланок. Проект геному людини – міжнародний проект наукових досліджень. Поряд з генами людини проект досліджував послідовності геномів і інших організмів, таких як Escherichia coli, плодова мушка, домашня миша та інших, що сприяло розумінню функціонування генів людини. Встановлення повної генетичної карти генів людини — важливий крок в розвитку медицини і інших аспектів охорони здоров'я.

У 1988 році Об’єднаний комітет (Національний інститут здоров’я та Міністерство енергетики США) представили проект “Геном людини”, у 1990 році була створена Міжнародна організація по вивченню геному людини (HUGO). Перший робочий проект геному був опублікований в 2000 році і завершений в 2003 році, подальший аналіз генів та секвенування окремих ділянок ДНК все ще продовжується.

Паралельний проект, розпочатий приватною американською компанією Celera Genomics (укр. Селера Джіномікс), проводився паралельно з міжнародним академічним проектом, і у 2000 році J. Craig Venter – власник

“Celera Genomics” – заявив про завершення його компанією розшифровки геному людини.

Безсумнівно, що успіху у розшифровці геному людини, а також швидкому розвитку молекулярної біології та генетики сприяли досягнення у сумісних областях науки: розробка нових підходів в хімії, поява потужних фізичних засобів дослідження внутрішньої будови речовини (рентгеноструктурний аналіз, електронна та атомна силова мікроскопія, методи, які базуються на використанні явища ядерного магнітного резонансу). Так нині, вже частково, чи повністю просеквеновано еукаріотичні геноми біля 200 видів живих істот, серед них (таблиця. 1.1) геноми людини, шимпанзе, миші, пацюка, кішки, собаки, курки, риби, дрозофіли, малярійного комара, черв’яків і дріжджів, біля десятка геномів рослин та інших організмів. Окрім того, просеквеновано більше тисячі бактеріальних геномів та біля сотні геномів архей. Встановлено, що величина генома людини сягає більше ніж 3 мільярди пар нуклеотидів (понад 30 тисяч генів), що з врахуванням інформації, яка одержана при його розшифровці, становить більше десяти терробайт даних.

З вищезазначеного слідує, що першим з головних факторів, що стимулював розвиток біоінформатики, стала необхідність ефективного опрацювання величезних масивів інформації, отриманих експериментальними методами молекулярної біології.

Організм	Повністю просеквеновано	Частково просеквеновано
Прокаріоти	840	662
Археї	56	8
Бактерії	784	655
Еукаріоти	22	175
Тварини	4	77
Ссавці	2	29
Птахи		2
Риби		8
Комахи	1	20
Плоскі черв'яки		1
Круглі черв'яки	1	7
Амфібії		
Рептилії		
Інші		11
Рослини	2	8
Наземні рослини	2	6
Зелені водорості		2
Грибки	10	64
Аскоміцети	8	51
Базидіоміцети	1	10
Інші грибки	1	3
Найпростіші	6	24
Apicomplexans	1	11
Kinetoplasts	1	4
Інші найпростіші	4	9
Всього:	862	837

Таблиця 1.1. Кількісь частково або повністю просеквенованих геномів.

Наприкінці XX століття вчені приступили до інвентаризації білків – основних функціональних структур організму і в середині 90-х років в молекулярній біології виник новий розділ – протеоміка. Протеоміка – наука, яка вивчає якісний і кількісний склад білків, що синтезуються клітиною. Так, порівняння протеом різних клітин в нормі і при патології дозволяє визначити механізми, які впливають на розвиток патологічних реакцій. Результатом цього є розробка методів медичної діагностики, що базуються на змінах

фізіологічного стану клітини і виражаються в пригніченні або стимуляції синтезу окремих білкових компонент клітини. В in vitro вже розроблені методи, які в змозі тимчасово або зовсім „виключити окремий ген”. Ультрачутливість і специфічність сучасного протеомного аналізу дозволяють реєструвати розвиток патологічних процесів на ранніх етапах захворювання, без його видимих симптомів, і ідентифікувати нові мішені для дії ліків, що створює основу пошуку нових патогенетичних терапевтичних засобів. На сьогодні одним з першорядних важливих медичних застосувань протеоміки є виявлення пухлинних клітин за їх білковими профілями, порівняння клітин до і після зазначених впливів. Такі можливості створюють нові перспективи як для діагностичної медицини, так і для фармацевтичної індустрії в плані створення нових лікарських препаратів.

Окремий розділ протеоміки – функціональна протеоміка – займається вивченням взаємодії різних білків у метаболічних шляхах, виявленням складу функціонально активних комплексів білків, що складають метаболічні ланцюги. Під терміном «функціональний геном» розуміють взаємодію біля двох тисяч генів за посередництвом експресованих ними білків, які на кожному етапі життєдіяльності організму забезпечують повноцінний клітинний цикл, проліферацію, диференціювання, клітинний гомеостаз та міжклітинну комунікацію.

Таким чином, другим фактором, що стимулював подальший розвиток біоінформатики, став розвиток самої молекулярної біології. Розвиток геноміки, протеоміки, функціональної геноміки і т.п. є неможливим без сучасних комп’ютерних методів таких, наприклад, як співставлення генетичних даних та на основі аналізу цих даних формування теоретичних припущень, які можуть бути перевірені дослідним шляхом та дають можливість отримувати нову інформацію. Так, досить часто існує можливість знайти схожий, вже вивчений білок, завдяки чому спрогнозувати функцію нового білку. У випадку, якщо схожі білки в базі даних відсутні, проводиться пошук подібних структур у різних білках. Наприклад, завдяки застосуванню співставлення нуклеотидних

послідовностей, було визначено, що вірус атипової пневмонії є продуктом злиття двох вірусів: людини і тварини, завдяки чому були визначені механізми його дії та методи нейтралізації.

Третім фактором стрімкого розвитку біоінформатики стало те, що на сьогодні біотехнологічні та медичні дослідження перемістилися з мікро- до наномасштабів та нанотехнологій, що знову таки, у значній мірі, пов'язано з розвитком геноміки, протеоміки та появою нових фізичних інструментів (наприклад, атомно-силовий мікроскоп) та методів роботи з найбільш важливими біологічними макромолекулами, які потрапляють в область наномасштабів – ДНК, РНК та білки. Для того, щоб контролювати ці біосистеми, необхідні інструменти, які безпосередньо взаємодіють з цими об'єктами у їх природному середовищі. Для розв'язання цих та багатьох інших задач нещодавно виникла нова галузь досліджень – нанобіотехнологія. Біотехнологія і нанотехнологія – це дві найбільш перспективні технології XXI сторіччя.

Очікується, що нанобіотехнологія з використанням методів біоінформатики зіграє важливу роль у генній терапії та інженерії, створенні високоселективних ліків, біомаркерів та біосенсорів, у наномедицині. Швидкий розвиток останньої, направлений на вирішення таких актуальних задач як діагностика, лікування та профілактика захворювань, розробка фармацевтичних засобів з метою збереження і покращення здоров'я людини, що неможливо без цілеспрямованого використання наноматеріалів у рамках молекулярних знань, щодо функціонування людського організму, а також сучасних комп'ютерних технологій та потужного апарату біоінформатики.

1.2. Предмет біоінформатики та її задачі

На сьогодні поки що немає єдиної точки зору, яким чином сформулювати предмет біоінформатики. В літературі (частіше – в російській) зустрічаються

визначення біоінформатики як дисципліни, що досліджує інформаційні процеси в живій природі, наприклад, біоінформатика – це «розділ інформатики, який враховує особливості збору, зберігання, обробки та використання інформації в біологічних системах» або біоінформатика – «напрям інформатики, який вивчає загальні закономірності та особливості інформаційних процесів в об'єктах біосфери». Автори цих визначень включають біоінформатику в групу інформаційних наук, а не біологічних. Між тим дослідження інформаційних аспектів існування живих систем слід віднести скоріше до теоретичної біології. До двох тільки що наведених визначень близьке ще одне: «Біоінформатика вивчає інформаційні процеси, які керують роботою клітини, органа, організму; зв'язком організмів між собою та оточуючим середовищем. Інформація видобувається, передається, регіструється, накопичується та утилізується. ...Основним носієм біоінформації є електромагнітні хвилі… Біоінформатика – це міждисциплінарний розділ науки, який відноситься до загальної біології, біофізики, електродинаміки, теоретичної та медичної генетики, радіофізики та психології мислення».

Інша група визначень наближає біоінформатику до комп'ютерної біології. Ресурс http://encyclopedia.thefreedictionarycom/bioinformatics (електронна енциклопедія) містить наступне визначення предмету біоінформатики: «Використання математичних та інформаційних методів для вирішення біологічних проблем, в основному, за допомогою розробки або застосування комп'ютерних програм, а також математичних моделей».

Іноді поняття «біоінформатика» можна зустріти (особливо у англомовній літературі) у вузькому технологічному змісті: під нею мають на увазі програмні засоби, спеціально призначені для аналізу «біологічних текстів» – як в загальноприйнятому розумінні слова «текст», так і у розумінні нуклеотидних і амінокислотних послідовностей. Наведемо приклад вузького розуміння дисципліни як інструмента біологічного дослідження: «Біологи починають використовувати засоби біоінформатики у лабораторній роботі так же, як у 1980 роки спеціалізовані методи молекулярної біології включалися в рутинні

біомедичні дослідження, а для складних процедур розроблялись стандартні набори».

Найбільш строгим визначенням біоінформатики на сьогодні представляється наступне: біоінформатика – область науки, що розробляє і застосовує технології інформатики для аналізу, систематизації молекулярно-біологічних даних, використовує фізико-математичні методи для моделювання процесів, що відбуваються на молекулярному рівні з метою виявлення структур, функцій та взаємодії макромолекул (ДНК, РНК, білків) з подальшим використанням цих знань при створенні нових лікарських препаратів та нановиробів для діагностики і лікування, а також отримання організмів з наперед заданими властивостями.

Основним принципом біоінформатики можна вважати те, що молекули нуклеїнових кислот та білків описуються за допомогою послідовності алфавітних знаків, причому число цих знаків обмежене (4 – для нуклеїнових кислот і 20 для опису білків). З точки зору лише інформатики, геном або білок – це довгий текст, що містить, наприклад, мільярд букв. Біоінформатика ж відрізняється тим, що її задачею є аналіз змісту цього тексту.

Найголовніше призначення біоінформатики – це розширення і поглиблення розуміння людиною біологічних процесів. Традиційна біологія з її підходами продовжують залишатися важливими, але для більш глибокого розуміння біологічних процесів необхідно посилити кількісну сторону молекулярної біології з акцентуванням уваги на поведінці різних систем організма і особливостей взаємовідношень між ними , що дозволить зрозуміти основи функціонування цілостного організму.

Зважаючи на великі обсяги інформації, що обробляється, розвиток і функціональність біоінформатики багато в чому залежать від прогресу у галузі розробки програмного забезпечення і апаратних засобів, тому в цьому плані виділяють наступні цілі біоінформатики:

- Розвиток програмних засобів і інформаційних ресурсів, які допомагають керувати даними, класифікувати та обробляти їх.

- Розробка нових, більш досконалих алгоритмів для роботи з великою кількістю даних.
- Організація даних таким чином, щоб дослідники могли мати доступ до інформації, що зберігається у біоінформаційних БД, а також могли заносити у ці бази нові записи про об'єкти, що ними досліджуються.
- Застосовування програмних засобів для аналізу даних і інтерпретації отриманих результатів таким чином, щоб вони мали біологічний зміст.

Перед біоінформатикою стоїть ряд важливих задач, викликаних сучасними потребами медицини і біотехнології:

- Знаходження генів у послідовностях ДНК різних організмів.
- Розвиток методів вивчення структури та (або) функцій нових розшифрованих послідовностей і відповідних структурних областей РНК.
- Визначення сімейств родинних послідовностей і побудова моделей еволюції.
- Вирівнювання подібних послідовностей і відновлення філогенетичних дерев з метою визначення еволюційних зв'язків.
- Збірка геномів із розшифрованих фрагментів послідовностей.
- Передбачення структури та функції білків.
- Знаходження мішеней та перспективних сполук для медикаментозної дії.
- Оцінка ролі окремих ділянок послідовності (мотивів, доменів) у функціонуванні білка. Розпізнавання таких ділянок розшифровує увесь спектр функцій конкретного білка.
- Побудова молекулярних моделей білків та дослідження механізму функціонування макромолекул, спираючись на їх моделі.
- Комп'ютерене конструювання ліків.

Біоінформатика займається експертними системами та системами інтелектуального аналізу даних (*data mining, knowledge discovery*) в біології. Нагадаємо про порівняльні особливості цих двох видів систем штучного

інтелекту. Експертна система є кінцевою; з неї видобувається знання, яке вже було закладене раніше експертом в базу знань, нове знання не створюється. Система інтелектуального аналізу даних – відкрита; при роботі з нею відбувається створенням нового знання, яке не було закладено в систему спочатку.

Область застосування методів штучного інтелекту в біології не обмежується молекулярним рівнем. Активно розвиваються нейроінформатика, еволюційна інформатика, обчислювальна екологія, обчислювальна популяційна біологія, комп'ютерне дослідження біологічного різноманіття – напрямки, що вивчають біологічну організацію на рівні клітин, тканин та органів, організмів та їх об'єднань. Всі ці напрямки охоплюють більш широке коло проблем комп'ютерної біології, але мають перетини і з біоінформатикою – там, де мова йде про аналіз великих масивів даних.

1.3 Методи біоінформатики

Завдяки стрімкому науково-технічному прогресу у галузі молекулярної біології, сучасним науковцям, як уже відмічалося раніше, доводиться працювати з колосальними обсягами експериментальних даних, аналіз яких неможливий без застосування сучасних інформаційних технологій і ефективних методів аналізу даних. У зв'язку з цим з'явилася необхідність створення математичного апарату для вирішення задач, особливостями яких є:

- необхідність роботи з великою кількістю даних;
- складні розрахунки, що часто потребують багато часу для виконання комп'ютером;
- особливі вимоги до продуктивності обчислювальної техніки.

Біоінформатика використовує методи прикладної математики, статистики, теорії ймовірності, інформатики та евристичні методи. Основні положення теорії ймовірностей, які необхідні біоінформатику ми розкриємо в останньому розділі підручника.

До математичних методів біоінформатики належать:

- алгоритми динамічного програмування (вирівнювання послідовностей, передбачення вторинної структури РНК і т.д.);
- ланцюги Маркова та приховані марківські моделі;
- трансформаційні граматики (використання методів аналізу мов для аналізу послідовностей);
- генетичні алгоритми – пошуковий метод, який засновано на селекції елементів в популяції (подібно еволюційній теорії Дарвіна);
- паралельні обчислення, пов’язані з використанням суперкомп’ютерів.

Щодо евристичних методів, то у літературі вони характеризуються як комплекс процедур, які спрямовані на скорочення перебору варіантів. Такі методи збільшують імовірність отримання придатного, але не завжди оптимального розв’язку певної задачі, що виникає, наприклад, через недостатню розробленість конкретної теорії, неповноту даних, або як це характерно для біоінформатики – через неоднозначність вихідних даних. Евристичні методи здатні знаходити розв’язки навіть у дуже складних, непередбачуваних ситуаціях. Але, слід пам’ятати, що не в усіх випадках евристичні методи можуть бути ефективними, за їх допомогою можуть бути знайдені наближенні розв’язки і тому такі методи часто поступаються точним алгоритмічним підходам. У загальному розумінні евристика визначає метод поведінки, що допомагає досягненню мети, але не може бути чітко охарактеризований. Незважаючи на вказані недоліки, евристичні методи дозволяють суттєво економити час обробки інформації, водночас, такі методи досить добре себе зарекомендували у біоінформатиці та широко використовуються на практиці у даний час.

У біоінформатиці евристичні методи використовуються для порівняння послідовностей, вони базуються на методах машинного програмування, у яких розв’язки знаходяться по встановленим дослідним шляхом правилам, а потім використовується обернений зв'язок для уточнення результату.

1.4 Статус біоінформатики

Біоінформатика отримала визнання не лише як область наукових досліджень, але і як учбова дисципліна.

Показником статусу признаної самостійної дисципліни вважається сукупність ознак:

видання спеціальної літератури – монографій, профільних журналів;

існування професіональних об'єднань і міжнародних асоціацій;

проведення наукових форумів;

розробка спеціальних учбових програм та наявність мережі спеціальних учбових закладів;

захист дипломів та дисертацій.

Розглянемо більш докладно ці ознаки.

Видання спеціальної літератури. Як було зазначено раніше, біоінформатика обє'днує широке коло знань у області біології, хімії, фізики, математики та інформатики. У зв'язку з цим, у друкованих періодичних виданнях питання біоінформатики розглядаються у рубриках видань різних наукових напрямків.

В області молекулярної біології і, зокрема, біоінформатики існує багато фахових видань та молекулярно-біологічних баз даних, що надають користувачеві через Інтернет вільний доступ не тільки до повних текстів наукових публікацій, а й до самих досліджуваних об'єктів – кодів нуклеотидних та амінокислотних послідовностей.

У природничих науках перше місце серед джерел опублікованої інформації займають журнали. Саме на цей вид наукової літератури частіше всього спостерігається попит серед спеціалістів по біології та медицині. Медико-біологічні науки традиційно є лідерами серед інших галузей по кількості журналів.

Найбільш вичерпним каталогом журналів вважається «Міжнародна бібліографія періодичних видань», яка названа на честь Констанції Ульріх: «Ulrich's International Periodicals Directory». Каталог (довідник) «Ульріх» видається з 1932 року, зараз видавцем є R.R. Bowker; Reed Elsevier Inc. У довіднику наводяться не лише періодичні видання, які публікують оригінальні статі, але також і реферативні та бібліографічні видання. Кількість включених найменувань неперервно зростає: 13-те видання каталогу «Ульріх» (1969-1970 роки) нараховувало 40 тис. найменувань, 21-е (1982 рік) – 63 тис., 30-е (1995 рік) – 120 тис., 39-е (2001 рік) – 164 тис. З видань, які включаються в каталог 40 % відносяться до точних, природничих та прикладних наук.

В останні роки каталог надається у інтерактивному режимі (http://www.ulrichsweb.com) та щоквартально на оптичних дисках.

Розділ «Біологія» в каталозі «Ульріх» містить 16 рубрик, серед яких виділено таку рубрику як використання комп'ютерів, до якої входять 125 періодичних видань.

Розділ «Медичні науки» каталогу «Ульріх» містить 32 рубрики, серед яких виділено таку рубрику як використання комп'ютерів, до якої входять 137 періодичних видань.

На початку 60-х років на ринку інформаційних послуг з'явився новий продукт. Це Вказівник цитованої літератури («Science Citation Index»). Цей Вказівник створив Ю. Гарфілд, який в 1958 році заснував Інститут наукової інформації США (ISI). На відміну від довідника «Ульріх» Вказівник містить лише журнали, що публікують оригінальні статі. В 1963 році перелік охоплював біля 600 найменувань журналів, а вже в 1985 році – біля 3300, на сьогодні цей перелік складає більше 7000 найменувань. В цьому переліку біологічні та медичні науки не об'єднані в спеціальні розділи, а представлені в загальному рубрикаторі (Subject Categories) самостійними рубриками.

Статті для вченого – це не лише засіб для того, щоб сповістити про свої нові ідеї і дослідження, але й спосіб заявити про свій пріоритет при отриманні нових знань. Пріоритет вченого підтверджується тим, що інші автори

посилаються на його роботи. Для характеристики значущості публікації для подальших досліджень використовується декілька параметрів. Розглянемо їх більш детально.

З існуючих параметрів найбільш популярним та широко розповсюдженим є імпакт-фактор (impact-factor). Поняття імпакт-фактору було запропоновано Ю.Гарфілдом в 1955 році і набуло широкого розповсюдження з 1963 року. Імпакт-фактор – чисельний показник важливості наукового журналу, він розраховується кожного року Інститутом наукової інформації. Розрахунок імпакт-фактора оснований на трирічному періоді, наприклад імпакт-фактор журналу в 2008 році розраховується за формулою:

$$I_{2008} = \frac{A}{B},$$

A – число посилань протягом 2008 року у журналах, які відслідковуються Інститутом наукової інформації (включаючи і журнал, імпакт-фактор якого розраховується) на статті, які були надруковані у журналі за 2006 – 2007 роках.

B – число статтей, надрукованих у даному журналі за 2006 – 2007 роки.

При розрахунках Інститут наукової інформації не враховує деякі типи публікацій (повідомлення, листи та ін.), крім того для нових журналів імпакт-фактор іноді розраховується лише для дворічних періодів. В таблиці 1.2 наведено класифікаційну шкалу рейтингу наукових журналів, а в таблиці 1.3 – імпакт-фактори 20 найбільш цитованих журналів в світі.

Імпакт-фактор	Рейтинг журналу
Більше 10	Дуже високий
5 - 10	Високий
1 - 5	Середній
0,5 – 1	Низький
0 – 0,5	Дуже низький

Таблиця 1.2. Класифікаційна шкала рейтингів наукових журналів

Ранг	Назва журналу	Кількість посилань	Імпакт-фактор	Кількість статей
1	Annual Review of Immunology	13 709	54,455	26
2	Annual Review of Biochemistry	16 591	36,278	28
3	CA-A Cancer Journal for Clinicians	3096	32,886	12
4	New England Journal of Medicine	143 124	31,736	378
5	Nature	326 546	30,432	889
6	Nature Medicine	31 696	28,740	137
7	Nature Immunology	6 297	27,868	134
8	Cell	139 765	27,254	350
9	Nature Genetics	44 050	26,711	222
10	Science	296 080	26,682	987
11	Pharmacological Reviews	6 854	26,568	23
12	Physiological Reviews	13 016	26,532	28
13	Nature Reviews Molecular Cell Biology	3 226	26,170	85
14	Annual Review of Neuroscience	7 348	24,091	19
15	Natural Reviews Neuroscience	2 317	24,047	86
16	Reviews of Modern Physics	15 338	23,672	32
17	Annual Review of Cell and Developmental Biology	6 482	22,870	24
18	Natural Reviews Genetics	2 052	21,762	83
19	Endocrine Reviews	9 219	21,643	40
20	Chemical Reviews	35 148	20,993	146

Таблиця 1.3. Імпакт-фактори 20 найбільш цитованих журналів в світі

У таблиці 1.4 для порівняння наведені імпакт-фактори російських та українських журналів.

Назва журналу	Імпакт-фактор		Назва журналу	Імпакт-фактор

1. Biochemistry (Moscow)	2001	2006
	0,762	1,368
2. Russian Journal of Bioorganic Chemistry (Російський журнал біоорганічної хімії)	–	–
3. Microbiology (Мікробіологія)	0,550	0,543
4. Applied Biochemistry and Microbiology (Прикладна біохімія і мікробіологія)	0,248	0,444
5. Biophysics	–	0,435
6. Russian Journal of General Chemistry	0,466	0,374
7. Molecular Biology (Молекулярна біологія)	0,490	0,330
8. Russian Journal of Genetics	0,054	0,254
9. Bulletin of Experimental and Biological Medicine	0,149	0,190
10. Molecular Genetics, Microbiology and Virology (Молекулярная генетика, микробиология и вирусология)	–	–

а)

	2001	2006
1.Experimental Oncology	0,269	–
2. Neurophysiology	0,087	–
3.Ukrainica Bioorganica Acta	–	–
4. Ukrainian Biochemical Journal (Український біохімічний журнал)	–	–
5. Cytology and Genetics (Цитологія і генетика)	–	–
6. Biopolymers and Cell (Біополімери і клітина)	–	–

б)

Таблиця 1.4 Імпакт-фактори а) російських, б) українських журналів.

З інших параметрів, що характеризують цінність публікацій, можна назвати наступні: період напівжиття та оперативність. Період напівжиття – це час протягом якого було опубліковано половину всій літератури по даній галузі, яка використовується на даний час. Цей термін було запропоновано в 1960 році американськими вченими Р. Бартоном та Р. Кеблером по аналогії з мірою швидкості розпаду радіоактивних речовин. Міра використання літератури визначається частіше всього за кількістю посилань на неї, тобто за бібліографічними посиланнями. На основі аналізу старіння літератури по будь-якій дисципліні можна зробити висновок про порівняльну цінність літератури,

що публікується, в майбутньому. Оперативність – міра публікації в журналі актуальних цьому журналу статей. Індекс оперативності (*immediacy index*) розраховується як відношення числа отриманих журналом у певному році посилань на статті, що були надруковані у журналі у тому же році до загальної кількості статей, які були надруковані у данному році у журналі.

Розглянемо, які основні журнали по проблемам біологічної інформації та застосувань комп'ютерної техніки в біології та медицині, випускаються сьогодні (таблиця 1.5).

Назва	Видається з	Імпакт-фактор	Основний зміст
Bioinformatics	1985 р.	6,019	Огляди програмних засобів та нових комп'ютерних розробок для біологів
The Bioinformer	1997 р.		Повідомлення про дослідження та найбільш важливі симпозіуми і конференції по біоінформатиці
Biological Cibernetics	1961 р.	1,474	Комунікації та керування в організмах та автоматах – експериментальні та теоретичні повідомлення; кількісні фізіологічні дослідження сенсорних органів і нервової системи, комп'ютерні дослідження процесів обробки сенсорної інформації.
Binary: Computing in Microbiology	1989 р.		Міжнародні матеріали по всім аспектам застосування обчислювальної техніки в мікробіології.
Briefings in Bioinformatics	2000 р.	24,370	Оглядові матеріали по базам даних та аналітичним засобам в молекулярній біології і генетиці, включаючи моделювання біологічних процесів,

			функціональну геноміку і протеоміку, фармакогеномику, статистичну генетику і генетичну епідеміологію, популяційну генетику людини.
Computer Methods and Programs in Biomedicine	1970 р.	0,624	Міжнародні матеріали по розробці та застосуванню комп'ютерних методів і програмних засобів в біомедичних дослідженнях та медичній практиці.
Journal of Biomedical Informatics	1969 р.	2,346	Використання комп'ютерів у біомедичних дослідженнях.
Computers and Medicine	1972 р.		Сучасні комп'ютерні технології в медицині, освіті.
Computers in Biology and Medicine	1971 р.	1,068	Міжнародні експериментальні та теоретичні дослідження по всім аспектам застосування комп'ютерів у біології та медицині.
In Silico Biology	1998 р.		Міжнародні дослідження по молекулярній комп'ютерній біології.

Таблиця 1.5. Журнали по проблемам застосувань комп'ютерної техніки в біології та медицині.

Ні в каталозі «Ульріх», ні в «Science Citation Index» ці видання не виділено у рубрику «Інформаційна біологія», «Комп'ютерна біологія», «Біоінформатика» або «Медична інформатика». В каталозі «Ульріх» інформація про відповідні журнали знаходиться у рубриках «Застосування комп'ютерів», які в розділах «Біологія» та «Медичні науки» нараховує 25 та 137 видань, відповідно. Крім того, журнали розсіяні по конкретним рубрикам розділів «Медичні науки», «Біологія» («Біотехнологія», «Мікробіологія»), «Хімія» («Застосування комп'ютерів»), «Комп'ютери» («Штучний інтелект»). В переліку для «Science Citation Index» такі журнали (зокрема, «Bioinformatics») знаходяться в трьох

рубриках: «Біологія» — різне; «Методи біомедичних досліджень»; «Комп'ютерна наука». До найбільш популярних журналів у даній галузі відносяться наступні.

American Medical Informatics Assocation Journal (JAMIA). ISSN 1067-5027. Початок видання: 1994 рік. Мова: англійська. Періодичність: 1 випуск у 2 місяці. Зміст: застосування комп'ютерних методів в охороні здоров'я, медичних дослідженнях та медичній освіті. Відображається у основних реферативних та бібліографічних джерелах: Curr. Cont. (Current Contents), Ind. Med. (Index Medicus), Ind. Sci. Rev. (Index to Scientific Reviews), Inpharma, Sci. Cit. Ind. (Science Citation Index). Видавник: American Medical Informatics Association, http://www.jamia.org/.

Bioinformatics. ISSN 1367-4803. Початок видання: 1985 рік. Мова: англійська. Періодичність: 1 випуск у місяць. Тип документів: наукові та освітні. Відображається: Biological Abstracts (Biol. Abstr.), Chemical Abstracts (Chem. Abstr.), Computer Abstracts (Comput. Abstr.), Curr. Adv. Ecol. Sci. (Current Advances in Ecological and Environmental Sciences), Curr. Biotech. Abstr. (Current Biotechnology Abstracts), Curr. Cont., Excerp. Med. (Excerpta Medica), Ind. Med., Ind. Sci. Rev., Inpharma, Sci. Cit. Ind. Видавник: Oxford Univ. Press, Academic Division, Oxford OX2 6DP, United Kingdom, http://www.oup.co.uk/journals.

The Bioinformer. ISSN 1462-1355. Початок видання: 1997 рік. Мова: англійська. Періодичність: 1 випуск в квартал. Тип документів: бюлетені новин. Видавник: European Bioinformatics Institute (EBI), Hinxton, Cambs CB10 1SD, United Kingdom, http://bioinformer.ebi.ac.uk/.

Biological Cybernetics. ISSN 0340-1200. Початок видання: 1961 рік Мова: англійська. Періодичність: 1 випуск в місяць. Тип документів: наукові та освітні. Відображається: Biol. Abstr., Chem. Abstr., Comput. Abstr., Curr. Adv. Ecol. Sci., Curr. Biotech. Abstr., Curr. Cont., Excerp. Med., Ind. Med., Ind. Sci. Rev., Inpharma, Neurosci. Abstr. (Neurosciences Abstracts), Sci. Cit. Ind. Видавник: Springer-Verlag, Heidelberg 69121, Germany, http://link.springer.de/link/service/journals/00422/index.htm.

Існування професіональних об'єднань і міжнародних асоціацій. За останні два десятитдіття в світі створено сотні біоінформаційних центрів (в США, Європі, Японії, Росії та ін.). Дослідженнями у галузі біоінформатики займається велика кількість державних та приватних установ. До таких організацій, можна віднести закордонні та українські інститути та центри:

- National Centre for Biotechnology Information, NCBI (USA), Національний центр біотехнологічної інформації – найбільша біологічна база даних (молекулярна біологія, генетика та біохімія) GenBank;
- European Bioinformatics Institute, EBI (UK), Європейський біоінформаційний інститут – база даних нуклеотидних послідовностей EMBL (European Molecular Biology Laboratory);
- National Institutes of Health, NIH (USA), Національний інститут здоров'я, що розробляє біоінформаційне програмне забезпечення та бази даних для біомедичних досліджень; вивчає характеристики ракових клітин с точки зору властивостей ДНК, іРНК, білкового, функціонального та фармакологічного рівня для пошуку нових засобів лікування раку;
- The National Science Foundation, NSF (USA), Національний науковий фонд. Підрозділ молекулярної та клітинної біології проводить фундаментальні дослідження процесів, які відбуваються в живих організмах, на молекулярному, субмолекуляному та клітинному рівнях;
- Інститут цитології та генетики СВ РАН, лабораторія теоретичної генетики, Новосибірськ, що забезпечує теоретичні та інформаційно-комп'ютерні основи геномних досліджень, генетики та селекції, молекулярної генетики та біології, генетичної та білкової інженерії, біотехнології, медичної генетики, генодіагностики, генотерапії;
- Інститут математичних проблем біології РАН, об'єднаний центр обчислювальної біології та біоінформатики, Московська обл., що

займається структурною та порівняльною геномікою, протеомікою, дослідженням основних молекляно-генетичних систем керування;

- Державний науково-дослідницький інститут генетики та селекції промислових мікроорганізмів, лабораторія біоінформатики, Москва, що займається створенням ліків нового покоління на основі рослинних токсинів, моноклональних антитіл; розробкою технологій отримання рекомбінантних білків людини, ферментів, антибіотиків, вітамінів, біокаталізаторів;
- Інститут молекулярної біології і генетики НАН України, відділ білкової інженерії і біоінформатики, Київ. Відділ створено для розвитку нових напрямків молекулярної біології, таких як білкова інженерія та біоінформатика;
- Інститут клітинної біології і генетичної інженерії НАН України, відділ генетичної інженерії, Київ. Відділ створено для розробки нових методів переносу генетичного матеріалу.

Центральну роль біоінформатики в сучасних біомедичних дослідженнях визнають такі провідні організації США, як Національний інститут здоров'я (NIH) та Національний науковий фонд (NSF). Так, в 1999 г. робоча група по біомедичним обчисленням NIH прийняла Ініціативний проект по біомедичній інформаційній науці та технологіям, який включає чотири основні програми:

(1) Програма забезпечення якості біомедичних досліджень, яка передбачає створення приблизно 20 центрів біоінформатики в університетах та незалежних науково-дослідних інститутах США. Ця програма направлена також на розвиток міждисциплінарної освіти з метою підвищити рівень комп'ютерної підготовленості спеціалістів по біології та медицині.

(2) Програма зберігання, оновлення, аналізу та видобування інформації.

(3) Програма адекватного фінансування на основі принципу надання грантів в розпорядження керівників дослідницьких груп.

(4) Створення національної комп'ютерної інфраструктури з можливостями розширення по мірі надходження нових даних.

Варто відзначити, що на сьогодні існує велика кількість біоінформаційних компаній, які займаються розробкою програмного забезпечення для молекулярної біології та біотехнології, а також створенням ліків. Наведемо деякі провідні біоінформаційні компанії:

- Celera Genomics Group (Rockville and Alameda, USA). Компанія розробляє та виробляє інструментальну базу та програмне забезпечення молекулярної біотехнології, що використовується для аналізу ДНК та РНК, дослідження малих молекул та білків з подальшим впровадженням результатів досліджень в фармакологічне виробництво, а також створення біосенсорів для діагностики.
- Genentech (San Francisco, USA). Виробництво біотерапевтичних препаратів для лікування ракових захворювань.
- Serono (Geneva, Switzerland). Світовий лідер в галузі репродуктивного здоров'я. Продукти компанії широко використовуються в неврології, а також при дослідженні процесів метаболізму та росту клітин.
- Biogen Idec (Cambridge, USA). Компанія виробляє препарати, які використовуються в онкології, неврології, імунології та лікуванні серцевих захворювань.

Проведення наукових форумів. Не менш важливим показником статусу самостійної дисципліни є проведення наукових конференцій і форумів. З біоінформатики постійно з 80-х років минулого сторіччя проводяться міжнародні конференції, багато з яких організуються і фінансуються фармацевтичними та біотехнологічними фірмами.

Розробка спеціальних учбових програм та наявність мережі спеціальних учбових закладів. Коли мова йде про отримання освіти, виникають наступні питання: «Де можна вивчити біоінформатику?», «Як її вивчити?», «Яким чином перейти від спеціалізації з обчислювальної науки до спеціалізації по біології та

навпаки?». Спеціальність «біоінформатика» в провідних університетах світу включено в систему вищої освіти. В університетах та біоінформаційних центрах існують курси дистанційної підготовки. Зараз найбільш розповсюджений шлях отримання освіти зі спеціальності «біоінформатика» – це магістратура з біоінформатики для осіб, які вже мають диплом бакалавра по біологічним або комп'ютерним наукам. При цьому практика показує, що в магістратуру з біоінформатики частіше поступають біологи або фізики, ніж спеціалісти по обчислювальним технологіям. Необхідність у застосуванні біоінформатики у якості технології в біомедичних дослідженнях зараз настільки висока, що ті, хто вже мають диплом, не знаходять часу для отримання нової фундаментальної освіти, а навчаються на курсах (іноді дистанційних) без відриву від основної роботи.

Існує три категорії освітніх курсів з комп'ютерних аспектів біоінформатики:

1) використання програмних засобів,

2) створення комп'ютерних програм,

3) розробка алгоритмів та їх теоретичних основ.

В процесі навчання, а також у подальшій роботі, створення (з одного боку) та застосування комп'ютерних програм (з іншого) представляють собою два різних напрямки: підготовка спеціалістів-дослідників («miners») та підготовка технологів («engineers»). При цьому необхідність в дослідниках вище, ніж необхідність в технологах.

Сотні вищих навчальних закладів світу пропонують програми навчання з комп'ютерної біології/біоінформатики: W.M.Keck-Centre по комп'ютерній біології (Пенсильванія та Техас), Департамент молекулярної біотехнології (Університет Вашингтона, Сіетл), Стенфордський університет (Каліфорнія), MIT (Масачусетс), Університет Ватерлоо (Онтаріо, Канада), Університет Пастера (Франція); Інститут Вейцмана (Ізраїль) та багато інших.

Лідер університетської освіти з біоінформатики – Великобританія, Німеччина, США та Індія. З середини 1980-х років важливим джерелом освіти

з біоінформатики стала EMBnet — федерація центрів («вузлів») з біоінформатики у Європі та світі. Офіційний британський вузол EMBnet – це Medical Research Council's Human Genome Mapping Project Resource Centre (MRC HGMP-RC) у Хінкстоні (Genome Campus at Hinxton) біля Кембриджу. На тому ж сайті розміщено два інших вузла – European Bioinformatics Institute (EBI) і Wellcome Trust Sanger Institute (WTSI).

У Московському державному університеті ім. Ломоносова в 2002 році було створено факультет біоінженерії та біоінформатики (http://www.fbb.msu.ru). Факультет готує кадри по спеціальності «біоінженерія та біоінформатика» для науково-дослідницьких інститутів і університетів, медичних закладів, промислових підприємств (особливо фармацевтичних і біотехнологічних виробництв). Відкрита спеціальність «Біоінформатика» на факультеті молекулярної та біологічної фізики Московського фізико-технічного інституту. На факультеті природничих наук Новосибірського державного університету працює кафедра інформаційної біології. На сайті кафедри (http://www. bionet. nsc. ru/chair/ cib/ InformBiologiya/ InformBiologiya.html) у якості предмету інформаційної біології заявлено «дослідження біологічних систем на трьох рівнях їх організації (молекулярно-клітинному, організменому та популяційному, а також екосистемному). При цьому інформаційна біологія «забезпечує інформаційно-комп’ютерні та теоретичні основи генетики і селекції, молекулярної генетики та молекулярної біології, генетичної і білкової інженерії, біотехнології, медичної генетики, генодіагностики, генотерапії, екології».

В Національному технічному університеті України “Київський політехнічний інститут” в 2006 році було створено кафедру біоінформатики. Також у Львівському національному Університеті ім. І. Франка кафедра біофізики та біоінформатики готує спеціалістів, бакалаврів та магістрів з спеціальності "Біофізика та біоінформатика". Раніше кафедра називалася кафедра біофізики та математичних методів у біології.

Захист дипломів та дисертацій. Говорячи про статус самостійної дисципліни завжди необхідно брати до уваги таку ознаку як захист дипломів та дисертацій з даної дисципліни. На сьогодні, нажаль, в Україні на даний момент не спостерігається такої посиленої уваги до розвитку біоінформатики, як приміром, у США, Європейському союзі, Японії, Індії, Росії та інших країнах, але уже зараз до Переліку спеціальностей, за якими проводиться захист дисертацій на здобуття наукових ступенів кандидата наук і доктора наук, присудження наукових ступенів і присвоєння вчених звань, внесено спеціальність медична та біологічна інформатика і кібернетика – 05.13.09. А ось у Російській Федерації з 1995 року можна захищати дисертації на здобуття наукового ступеня кандидата або доктора наук зі спеціальності «Біоінформатика» (з 1995 року – по біологічним та фізико-математичним наукам, а з 2000 року – також по медичним та сільськогосподарським), оскільки ця дисципліна включена в «Номенклатуру спеціальностей наукових робітників» (шифр 03.00.28). Незважаючи на те, що науковий ступінь може присвоюватися з декількох галузей, включаючи фізико-математичні науки, спеціальність віднесено до розділу «Біологічні науки», а не до розділу «Інформатика, обчислювальна техніка та керування».

Наведемо основні області дослідження у відповідності з паспортом спеціальності 03.00.28 – «Біоінформатика»:

1) Дослідження еволюції живої природи за допомогою методів інформатики та математики.
2) Комп'ютерне та математичне моделювання інформаційних процесів в біологічних системах.
3) Комп'ютерна генетика: розшифровка та моделювання структурної організації генів та геномів, а також кодуємих генами білків; аналіз мутацій та ін.
4) Комп'ютерна нейробіологія: моделювання природних нейронних систем, розробка нейромереж та ін.

5) Дослідження екологічних систем за допомогою інформаційних технологій.
6) Комп'ютерне моделювання біологічної дії ксенобіотиків.
7) Комп'ютерне моделювання процесів отримання, накопичення, обробки та систематизації біологічних та медичних даних.
8) Комп'ютерне розпізнавання та синтез зображень біологічних об'єктів.
9) Створення нових інформаційних технологій на основі результатів досліджень живої природи.
10) Організація та використання автоматизованих банків даних з біології та медицини, в тому числі банків міждисциплінарних даних.
11) Розробка інтелектуальних систем аналізу та прогнозування властивостей біологічних об'єктів на основі спеціалізованих баз та банків даних.
12) Створення систем інформаційного забезпечення та підтримки біологічних та медичних досліджень.

Таким чином, підсумовуючи усі ознаки, що характеризують біоінформатику як науку і учбову дисципліну, можна стверджувати, що біоінформатика розвивається дуже стрімкими темпами і займає надзвичайно важливе місце у системі наук.

Запитання до розділу 1

1. Які події можна вважати початком бурхливого розвитку генетики та молекулярної біології?
2. Що Вам відомо про Проект геному людини?
3. Що таке протеоміка?
4. Назвіть основні чинники, що стимулювали виникнення та розвиток науки біоінформатика.
5. Сформулюйте визначення біоінформатики.
6. У чому полягають цілі біоінформатики?
7. Які задачі біоінформатики?

8. Які методи використовує біоінформатика?
9. Що таке евристичні методи?
10. Сукупність яких ознак вважається показником статусу признаної самостійної дисципліни?
11. Які каталоги журналів Вам відомі?
12. Що таке імпакт-фактор?
13. Які журнали, присвячені проблемам біологічної інформації та застосувань комп'ютерної техніки в біології та медицині Вам відомі?
14. Які навчальні заклади світу пропонують програми навчання по комп'ютерній біології/біоінформатиці?

Література до розділу 1

1. С. Игнасимуту, Основы биоинформатики, [пер. с англ. А.А. Чумичкина], М.-Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. – 320 с.
2. А.И. Арчаков Геномика, протеомика и биоинформатика — науки XXI столетия //Фармацевтический вестник №9 (208) 13 марта 2001 г
3. http://123bioinformatics.com/?page_id=261
4. http://anil.cchmc.org/University.html
5. Игнасимуту С**.** Основы биоинформатики. – М. – Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. – 320 с.
6. Глазко В.И., Глазко Г.В. Введение в генетику, биоинформатика, ДНК-технология, генная терапия, ДНК-экология, протеомика, метаболика – К.:КВІЦ, 2003.
7. Lesk M. Introduction to Bioinformatics – Oxford University Press Inc. New York, 2002.

8. Чехун В.Ф. Функціональний онкогеном — основа сучасної діагностики та нової стратегії- онкологія, том 8, №2, 2006.

РОЗДІЛ 2

ОБ'ЄКТИ ДОСЛІДЖЕННЯ БІОІНФОРМАТИКИ

Для того, щоб розібратися з широким колом питань, які вирішує біоінформатика, необхідно розглянути основні поняття молекулярної біології, тобто зрозуміти фундаментальні властивості та функції таких молекул як нуклеїнові кислоти та білки. Дані молекули і являються об'єктами дослідження біоінформатики.

Нуклеїнові кислоти – це важливі біополімери, які зберігають та передають генетичну інформацію у живих організмах. У живих клітинах міститься два типи нуклеїнових кислот – дезоксирибонуклеїнова кислота (ДНК) і рибонуклеїнова кислота (РНК).

2.1 Дезоксирибонуклеїнова кислота

У 1869 році швейцарський біохімік Фрідрих Мішер виявив у ядрі клітин сполуки, що мають кислотні властивості. Ці сполуки було названо нуклеїновими кислотами. Нуклеїнові кислоти були присутні у клітинах усіх організмів, починаючи з найпростіших, і закінчуючи вищими організмами. Хімічний склад, структура і основні властивості цих речовин були подібними для всіх досліджених організмів.

ДНК в організмі виконує дві функції: містить інформацію, на основі якої функціонує клітина, а також передає цю інформацію нащадкам.

Одним із перших доказів ролі ДНК у передачі спадкової інформації стали досліди з трансформації у бактерій пневмококів, одних із збудників запалення легенів. Трансформація у бактерій – це залучення ділянок ДНК бактерій одного штаму у ДНК іншого штаму і передача йому своїх властивостей. Відомо дві форми пневмококів: з полісахаридною капсулою і без такої капсули. Обидві ці ознаки спадкові. Капсульні пневмококи при зараженні ними мишей,

спричиняють запалення легенів, від якого миші гинуть. Безкапсульна форма для мишей нешкідлива.

У 1928 році англійський бактеріолог Ф. Гриффітс заражав мишей сумішшю, яка містила вбиті нагріванням капсульні пневмококи і живі пневмококи без капсул. Вчений вважав, що від такого зараження миші не захворіють на запалення легенів. Але після експерименту миші все ж таки загинули. Отже виявилося, що вбита форма певним чином передавала свої властивості живим клітинам безкапсульної форми. З'ясувати за допомогою якої речовини відбувається передача спадкової ознаки вдалося лише через 16 років у 1944 році. Американським ученим А. Евері, К. Мак-Леоду і М. Мак-Карті після ряду дослідів вдалося довести, що ліпіди і вуглеводи не мають ніякого відношення до передачі спадкових властивостей капсульного пневмококу, а цей процес опосередковується спіральними біологічними молекулами. Таким чином було з'ясовано, що спадкову інформацію у живих істотах зберігає і передає молекула ДНК, що і було підтверджено експериментально.

Молекула ДНК складається з двох полінуклеотидних ланцюгів. Кожний ланцюг закручений у спіраль. Обидва ланцюги сполучені перемичками, звиті разом, таким чином вони утворють подвійну спіраль.

Структурними одиницями такої спіралі є нуклеотиди. Кожний нуклеотид складається з трьох компонентів: 1) моносахарид з п'ятьма атомами вуглецю (дезоксирибоза, рис. 2.1); 2) залишок фосфорної кислоти; 3) сполук чотирьох видів, які містять азот і мають хімічні властивості основ (рис. 2.2). У ДНК ці чотири основи представлені цитозином (Cytosine, C), тиміном (Thymine, T), гуаніном (Guanine, G) та аденіном (Adenine, A). За своєю будовою аденін і гуанін є дициклічними молекулами, що належать до групи пуринів, а цитозин і тимін – моноциклічними молекулами і відносяться до піримідинів. Ці чотири типи нуклеотидів можуть з'єднуватися через фосфати у будь-якій послідовності і утворювати одномірний ланцюг. Саме чергування нуклеотидів містить інформацію про властивості живого організму.

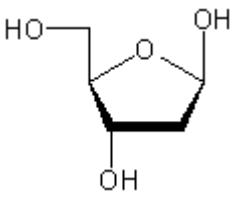


Рис. 2.1 Будова дезоксирибози

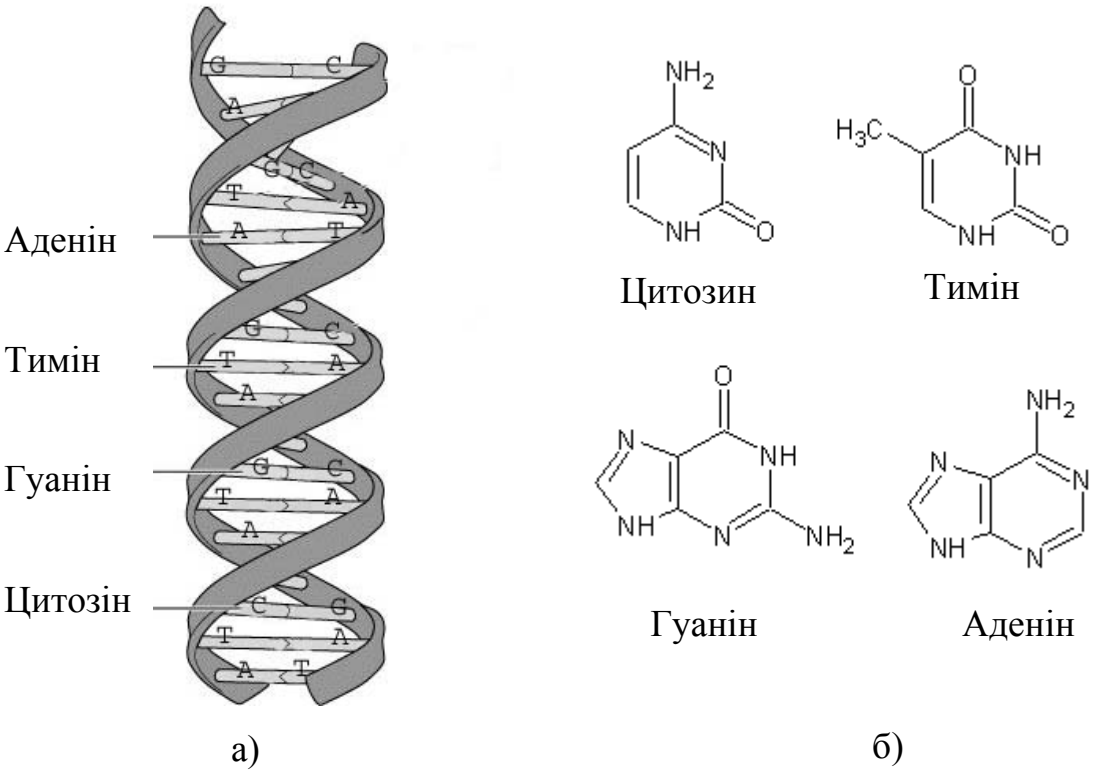


Рис.2.2. а) Ланцюг ДНК, б) будова азотистих основ

Нуклеотиди об’єднуються у ланцюг ковалентними зв’язками між дезоксирибозою одного і залишком фосфорної кислоти іншого нуклеотида. Ковалентний зв'язок – це хімічний зв'язок між двома атомами, який забезпечується спільною для них електронною парою.

Пригадаємо, що водневий зв'язок – це міжмолекулярний зв'язок, який існує завдяки перерозподілу зарядів; такий зв'язок, наприклад, можуть утворювати хлор, фтор, азот, або кисень з воднем. Водневі зв’язки утворюються між двома нуклеотидами у різних ланцюгах і за допомогою таких зв’язків два ланцюга

з’єднуються у одну подвійну спіраль. Хоча такий зв'язок і слабкий, однак мільярди таких зв’язків здатні утримати дві молекули разом. Основна властивість нуклеотидів проявляється у тому, що вони утворюють стабільні водневі зв’язки не з будь-яким нуклеотидом, а лише з одним із них, а саме:

A – сполучається лише з T;

C – сполучається лише з G.

Наприклад, якщо перша спіраль описується послідовністю: CGATT, то вона буде з’єднуватися зі спіраллю, що відповідає послідовності GCTAA. Така властивість називається комплементацією (complement – доповнення), а ланцюги – *комплементарними*.

Принцип комплементарності дозволяє зрозуміти механізм унікальної властивості молекули ДНК – їх здатність самовідтворюватися. ДНК – це єдина речовина у живих клітинах, що має подібні властивості. Процес самовідтворення молекул ДНК відбувається за активною участю ферментів. Так звані, розплітаючі білки послідовно розривають водневі зв’язки між ланцюгами ДНК. У результаті цього, утворюються одиночні ланцюги ДНК до яких за принципом комплементарності приєднуються вільні нуклеотиди, які завжди у достатній кількості знаходяться у ядрі. У ланцюгу, що утворився, виникають вуглеводно-фосфатні і водневі зв’язки. Таким чином, у процесі самовідтворення ДНК з однієї молекули синтезуються дві нові (див. рис. 2.3). Процес точного самовідтворення ДНК називається *реплікацією*. Окрім функції збереження спадкової інформації ДНК також «керує» процесами, що відбуваються у клітині, щоб їх зрозуміти, необхідно розглянути структуру і властивості РНК і білків.

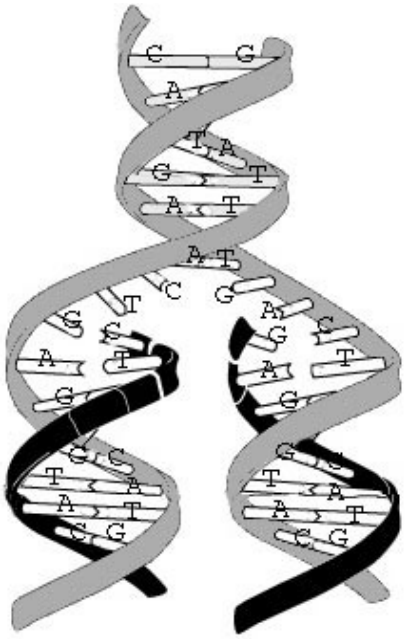
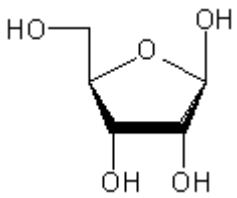


Рис 2.3. Реплікація ДНК

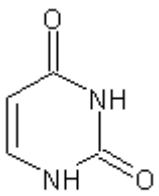
2.2 Рибонуклеїнова кислота

Рибонуклеїнова кислота (РНК), як і ДНК, містить залишок фосфорної кислоти, різниця полягає у тому, що РНК містить рибозу (ДНК містить дезоксирибозу) і азотисті основи – аденін, цитозин, гуанін і урацил (замість урацилу ДНК містить тимін). *Рибоза* – це моносахарид, що належить до класу пентоз (рис. 2.4, а). Наявність рибози значно змінює фізичні і хімічні властивості молекули РНК. У той час, як молекула ДНК дуже міцна, молекула РНК дуже нестійка і може легко розпадатися. Деякі вчені вважають, що першими полінуклеотидами, що з'явилися у ході біологічної еволюції були молекули РНК, і лише пізніше для довготривалого збереження генетичної інформації з'явилися хімічно більш стабільні молекули ДНК.

Що стосується азотистих основ то у РНК замість тиміну (Т) з'являється урацил (U), який комплементарний до аденіну (рис. 2.4, б), а також U може утворювати дуже слабкий зв'язок з G, у той час, коли у молекулі ДНК зв'язок G-T неможливий.



а)



Урацил

б)

Рис. 2.4. а) Будова рибози, б) будова урацилу.

Третя відмінність молекул РНК полягає у тому, що вони, зазвичай, значно коротші і можуть існувати у одноланцюговії формі, у той час, коли молекули ДНК існують у формі подвійної спіралі. РНК містяться головним чином в цитоплазмі клітин. Ці молекули синтезуються в клітинах всіх живих організмів, а також в РНК-вмісних вірусах.

За структурою розрізняють дволанцюгові і одноланцюгові РНК. Дволанцюгові РНК зберігають генетичну інформацію у деяких вірусів, тобто виконують у них функції хромосом. Одноланцюгових РНК існує декілька видів, які розрізняються, за місцем розташування у клітині, а також функціями, які вони виконують. Опишемо деякі з них.

Інформаційна РНК (іРНК) – РНК, яка слугує посередником при передачі інформації, яка закодована у ДНК, до рибосом, де синтезуються білки живого організму. Молекули іРНК можуть складатися з 300 – 30000 нуклеотидів.

Більшість РНК не кодують білок. На даний час дуже докладно вивчені такі типи некодуючих РНК, як рибосомні і транспортні РНК. Існують також класи РНК, що відповідальні за регуляцію генів; некодувальні РНК, що здатні каталізувати хімічні реакції, такі як розрізання і зшивання молекул РНК; по аналогії з білками, що каталізують хімічні реакції – ензими (ферменти), каталітичні молекули РНК називаються рибозимами.

Більшу частину цитоплазми (до 80-90%) складає *рибосомна РНК* (рРНК), що міститься у рибосомах, і слугує для них структурною та каталітичною

основою. Молекули рРНК відносно невеликі і складаються з 3-5 тис. нуклеотидів.

Транспортні РНК (тРНК) включають 76-85 нуклеотидів і виконують декілька функцій. Вони переносять специфічні амінокислоти у місце синтезу пептидного зв'язку у рибосомі, за принципом компліментарності вони «впізнають» триплет іРНК, який відповідає амінокислоті, що переноситься, і точно орієнтують амінокислоту на рибосомі.

2.3 Білки

Наступним, не менш важливим об'єктом дослідження біоінформатики є білки. Білки або протеїни (що у перекладі з грецької «перші» або «найважливіші») – це важливий клас біологічно активних речовин. Білки, які відіграють провідну роль у клітині, присутні у вигляді головних компонентів у будь-яких формах живої матерії та складають більше половини сухої ваги більшості організмів. Білки виконують ряд важливих функцій (рис.2.5). Вони слугують тими інструментами, завдяки яким генетична інформація отримує своє реальне втілення. Головна функція *білків-ферментів* – каталіз біохімічних реакцій, вони приймають участь у великій кількості перетворень, які відбуваються у живій клітині і складають основу її метаболізму. Особливе значення мають такі універсальні ферментні системи як ДНК-, та РНК-полімерази.

Рис. 2.5. Біологічні функції білків

Транспортні білки плазми крові зв'язують і переносять специфічні молекули, або іони із одного органа у інший. Гемоглобін, що міститься у еритроцитах, при проходженні крові через легені зв'язує кисень і доставляє його до специфічних тканин, де кисень вивільнюється і використовується для окиснення компонентів харчування. Плазма крові містить ліпопротеїни, які здійснюють перенесення ліпідів із печінки у інші органи. У клітинних мембранах присутній ще один тип транспортних білків, який в змозі зв'язувати глюкозу, амінокислоти і інші речовини та переносити їх крізь мембрану всередину клітини.

Резервні білки знаходяться у насінні багатьох рослин. Найбільш відомими прикладами таких білків слугують білки насіння пшениці, кукурудзи та рису. До цих білків також відноситься яєчний альбумін – основний компонент яєчного білку і казеїн – головний білок молока.

Рухальні білки надають клітині здатності скорочуватися, змінювати форму або пересуватися. Актин і міозин функціонують у скорочувальній системі

скелетного м’язу, а також у багатьох не м’язових клітинах. Інший білок тубулін – білок з якого побудовані мікротрубочки, що входять до складу цитоскелету еукаріотів. Такі мікротрубочки є важливими елементами джгутиків і ворсинок за допомогою яких пересуваються клітини.

Структурні білки утворюють волокна, які навиті одне на одне або укладені пласким шаром; вони виконують опорну або захисну функцію, зміцнюючи біологічні структури. Головним компонентом хрящів і сухожиль є фібрилярний білок колаген, який має дуже велику міцність. Зв'язки, у свою чергу, побудовані з еластину – структурного білку, здатного розтягуватися у двох вимірах. Волосся, нігті і пір'я майже цілком складаються з міцного нерозчинного білку кератину. Структурний білок фіброїн є головним компонентом шовкових ниток і павутиння.

Захисні білки захищають організм від вторгнення інших організмів та від пошкоджень. Імуноглобуліни або антитіла утворюються у хребетних – це спеціалізовані білки, які виробляються у лімфоцитах. Такі білки мають здатність розрізняти бактерії, віруси або чужорідні білки, що потрапили у організм, а потім нейтралізувати їх. Фібриноген і тромбін – це білки, що приймають участь у процесі згортання крові, вони захищають організм від втрат крові при пошкодженні судинної системи.

Регуляторні білки беруть участь у системі регуляції клітинної і фізіологічної активності, до них належить більшість гормонів, такі як інсулін, що регулює обмін глюкози, гормон росту. Інші регуляторні білки, які називаються репресорами, регулюють біосинтез у бактеріальних клітинах.

Існує велика кількість інших білків, функції яких доволі незвичайні. Наприклад, білок монелін, який утворюється в одній з африканських рослин має дуже солодкий смак. Він став предметом дослідження як нетоксична речовина, що не сприяє ожирінню і може бути використаний у якості замінника цукру. Плазма крові деяких антарктичних риб містить білки з властивостями антифризу, що захищають кров таких риб від замерзання. «Шарніри» у місцях закріплення крил у деяких комах складаються із білка резиліна, який має майже

ідеальну еластичність. Однак, усі ці білки, що так сильно відрізняються один від одного своїми властивостями і функціями, побудовані з одних і тих самих 20 амінокислот (Таблиця 2.1.)!

Біологічна функція білків тісно пов'язана з їх трьохмірною структурою, тому невелика зміна їх структури приводить до змін у функціонуванні білка. На початку минулого століття Е. Фішером було встановлено, що за своєю структурою білок являє собою поліпептидний ланцюг, що складається з амінокислотних залишків. Кожний ланцюг розташований і згорнутий строго визначеним способом.

Існує двадцять основних видів амінокислотних залишків. Однак модифікація білка іноді збільшує різноманітність амінокислот. Окрім того, у деякі білки включаються різні кофактори – невеликі молекули, іони, цукри, нуклеотиди, фрагменти нуклеїнових кислот і т.п. Вони можуть приєднуватися до ділянок ланцюга за допомогою ковалентного зв'язку або приєднуватися за допомогою специфічних механізмів. Наприклад, такі ферменти як пероксидаза и каталаза містять у своєму складі такий кофактор як залізо, аскорбінатоксидаза – мідь, а алкогольдегідрогенеаза – цинк, без цих кофакторів ферменти не активні і не виконують своїх функцій.

Нековалентні взаємодії, які підтримують просторову будову білка значно слабші ніж хімічні зв'язки, за допомогою яких з'єднані амінокислоти у білковому ланцюгу. Послідовність амінокислот називається «первинною структурою білка».

За загальним типом будови можливо розрізнити три основних типи білків:

1. Фібрилярні білки – білки, які утворюють великі агрегати, їхня структура високо регулярна і утримується, переважно, завдяки взаємодіям між різними ланцюгами.
2. Мембранні білки – розташовані частково у мембрані та за її межами, їх частини виступають з мембрани і знаходяться у водному середовищі. Внутрішньомембранні частини таких білків, як і фібрилярні білки – високо регулярні, та їхня структура стабілізується водневими та

гідрофобними зв’язками, зрозуміло, що розмір таких регулярних частин обмежений товщиною мембрани.

Назва	Позначення	
	трьохсимвольне	односимвольне
Гліцин	Gly	G
Аланін	Ala	A
Валін	Val	V
Лейцин	Leu	L
Ізолейцин	Ile	I
Метіонін	Met	M
Серин	Ser	S
Треонін	Thr	T
Цистеїн	Cys	C
Фенілаланін	Phe	F
Тирозин	Tyr	Y
Триптофан	Trp	W
Аспартат	Asp	D
Глутамат	Glu	E
Аспарагін	Asn	N
Глутамін	Gln	Q
Гістедин	His	H
Лізин	Lys	K
Аргінін	Arg	R
Пролін	Pro	P

Таблиця 2.1. Основні амінокислоти, що входять до складу білка.

3. Водорозчинні білки знаходяться у водному середовищі, їхня структура є нерегулярною і вирізняється певним розмаїттям та підтримується дисперсійними силами водневими та гідрофобними зв’язками.

Зрозуміло, що наведений вище розподіл дещо неточний, тому що білок може складатися із фібрилярного «хвоста» і глобулярної головки (наприклад, таку будову має міозин) і т.п.

На даний час відомо сотні тисяч білкових амінокислотних послідовностей. Для зберігання амінокислотних послідовностей створені спеціальні банки даних, наприклад, такий як SwissProt, а для зберігання просторових структур білків створено банк білкових структур. Найбільш повні відомості у даний час існують про водорозчинні глобулярні білки, тому що водорозчинні білки легше виділяти у вигляді окремих молекул, а їх структуру легко вивчати у кристалах за допомогою рентгену, а у розчинах за допомогою ЯМР спектроскопії (ядерного магнітного резонансу). Для мембранних і фібрилярних білків розшифровані лише деякі просторові структури і окремі фрагменти.

2.4. Ген, генетичний код, геном.

Як було зазначено раніше, ланцюг ДНК являє собою нуклеотидну послідовність. Таку послідовність можна розділити на певну кількість ділянок, які кодують, або не кодують інформацію, яка використовується при синтезі білків. Такі ділянки ДНК і називаються генами.

Перше визначення гену звучало наступним чином «ген – це одиниця спадковості, один ген відповідає одній властивості», але це визначення, у певній мірі можна віднести до таких властивостей як забарвлення квітки, або форми насінини. Однак, у живих організмах існують властивості, які обумовлені декількома або навіть декількома десятками генів, наприклад, ріст організму. Для більшості випадків, один ген відповідає одному білку у

організмі, але існують і такі гени, які взагалі не кодують білок, або такі, які кодують одразу два, або більше білків.

Гени кодують не лише білки, але і РНК, тому у повній мірі ген можна визначити як безперервну ділянку ДНК, що містить інформацію, необхідну для побудови білка або молекули РНК. *Ген* – це фізична одиниця спадковості (або спадковий фактор за Менделем), упорядкована послідовність нуклеотидів ДНК, в якій закодовано інформацію про один або декілька продуктів гену. Продуктами гену можуть бути: іРНК(а потім білок), рРНК, тРНК і білки.

Один і той самий ген може мати різні форми, *алель* – це одна з форм одного гена. Різні алелі виникають у результаті мутацій або внутрішньої перебудови генів.

Варто зазначити, що ДНК еукаріот є переривчастими, вони складаються з кодуючих ділянок — екзонів, розділених некодуючими — інтронами. Кількість таких ділянок різна для різних генів, наприклад ген овальбуміну курей включає 7 інтронів, а ген проколагену ссавців – 50 інтронів. Ділянки ДНК, які відповідають екзонам, на відміну від інтронів, повністю представлені в молекулі інформаційної РНК, що кодує первинну структуру білка. ДНК прокаріот інтронів не мають. Окрім екзонів та інтронів ген також містить регуляторні райони, які контролюють його експресію. Кожний ген розташовується у визначеному (певному) місці – локусі хромосоми. Довжини генів мають різні значення від 190 до 16000 п. н. У клітині існують певні механізми, які здатні точно розпізнавати в послідовності ДНК положення початку й кінця кожного гена.

Так як білок являє собою ланцюг амінокислотних залишків, для його синтезу необхідно ідентифікувати кожну амінокислоту, що входить у його склад. Саме цю операцію й виконує молекула ДНК, використовуючи триплети нуклеотидів для кодування всіх амінокислот. Нуклеотидний триплет називають *кодоном*. У таблиці (2.2) наведено відповідності всіх можливих триплетів амінокислотам. Така відповідність називається *генетичним кодом*. У таблиці триплети нуклеотидів наведені з урахуванням складу послідовності РНК. Це

пов'язано з тим, що саме молекули РНК забезпечують зв'язок між ДНК і безпосереднім синтезом білка.

Перша позиція	Друга позиція				Третя позиція
	G	**A**	**C**	**U**	
G	Gly	Glu	Ala	Val	**G**
	Gly	Glu	Ala	Val	**A**
	Gly	Asp	Ala	Val	**C**
	Gly	Asp	Ala	Val	**U**
A	Arg	Lys	Thr	Met	**G**
	Arg	Lys	Thr	Ile	**A**
	Ser	Asn	Thr	Ile	**C**
	Ser	Asn	Thr	Ile	**U**
C	Arg	Gln	Pro	Leu	**G**
	Arg	Gln	Pro	Leu	**A**
	Arg	His	Pro	Leu	**C**
	Arg	His	Pro	Leu	**U**
U	Trp	Стоп	Ser	Leu	**G**
	Стоп	Стоп	Ser	Leu	**A**
	Cys	Tyr	Ser	Phe	**C**
	Cys	Tyr	Ser	Phe	**U**

Таблиця 2.2 Універсальний генетичний код

Як видно з таблиці 2.2 існує 64 можливих триплети нуклеотидів і 20 амінокислот, які кодуються за допомогою цих триплетів. Одній і тій же

амінокислоті може відповідати кілька триплетів. Наприклад, кодони AAG і AAA кодують лізин. Навпаки, три кодони в таблиці не кодують ніяку амінокислоту, а служать сигналом закінчення гена. Ці спеціальні кодони позначені в таблиці 2.2 словом «Стоп». Наведений у таблиці генетичний код, називається універсальним, тому що використовується переважною більшістю живих організмів, однак деякі організми використовують трохи видозмінений код.

Таким чином, генетичний код має характерні особливості:

1. Універсальність – код однаковий для усіх організмів. Один і той самий триплет (кодон) кодує одну і ту ж саму амінокислоту, виключення складають деякі генетичні системи (наприклад, гени мітохондрій і хлоропластів) для яких кодони відрізняються від стандартного коду, властивого більшості організмів.

2. Кожна ділянка ДНК містить інформацію не більш ніж про один білок, іншими словами, якщо ділянка ДНК кодує білок, то вона не може кодувати (починаючи з якого-небудь іншого нуклеотида) інший білок.

3. Виродженість – більшість амінокислот можуть кодуватися декількома кодонами. Виняток складають дві амінокислоти – метіонін і триптофан, які мають лише по одному варіанту кодона.

4. Між генами існують стоп-кодони – три спеціальних триплети (UAA, UAG, UGA), кожен з яких означає закінчення синтезу поліпептидного ланцюга.

5. Всередині генів стоп-кодонів немає.

Фундаментальний набір генетичної інформації називається *геномом* і складається з однієї або декількох молекул ДНК (у деяких вірусів РНК), які організовані у вигляді хромосом. У бактерій розміри генома не бувають менше ніж 0,5 мільйонів пар нуклеотидів (п.н.), а максимальний розмір геному бактерій біля 10 мільйонів п.н., у дріжджів – порядку 12 мільйонів п.н., у черв'я нематоди – 97 мільйонів п.н., а у людини – 3 мільярда п.н.. Число хромосом у геномі організму характеризує його приналежність до певного виду. Наприклад, кожна клітина людини містить 46 хромосом, у той час у миші це число рівне 40.

У таблиці 2.3. наведені набори хромосом і розміри геному деяких характерних видів.

Вид організму	Число хромосом	Розмір геному (пар нуклеотидів)
Sacchammyces cerevisiae (дріжджі)	32	$1{,}2 \cdot 10^7$
Caenorhabditis elegans (черв'як)	12	10^8
Drosophila melanogaster (плодова мушка)	8	$2 \cdot 10^8$
Homo sapiens (людина)	46	$3 \cdot 10^9$

Табл.иця 2.3. Розмір геному певного виду

У прокаріот аналогом хромосоми виступає єдина у клітині молекула ДНК. Окремо від хромосомної ДНК існують молекули ДНК, що називаються плазмідами, вони зазвичай круглі і дволанцюжкові. Плазміди частіше за все зустрічаються у бактерій. У деяких еукаріот також можна зустріти плазміди, наприклад у дріжджів *Saccharomyces cerevisiae*. У одній клітині може бути від одної копії (особливо для великих плазмід) до кількох сотень або навіть тисяч копій тієї ж плазміди (особливо для певних штучних плазмід).

У клітинах еукаріотів хромосоми можуть бути присутні як у одиничному, так і у парному вигляді. Клітини, що містять парні хромосоми, називають диплоїдними. Наприклад, геном людини складається з 23 пар хромосом. Дві хромосоми, що формують пару називаються гомологічними, а гени у одному члені пари відповідають генам у другому члені. Деякі гени абсолютно ідентичні у батьківській і материнській частині пари – наприклад, ген, що кодує гемоглобін (білок, що здійснює перенесення кисню у крові). Інші гени можуть бути присутніми у альтернативних формах, які отримали назву алелі.

Клітини, що містять лише по одному елементу з кожної пари хромосом, називають гаплоїдними. Ці клітини беруть участь у статевому розмноженні організмів. Коли гаплоїдна клітина матері зливається з галоїдною клітиною батька, утворюється яйцеклітина, яка знову буде диплоїдною. Гаплоїдні клітини формуються у процесі мейозу, при якому клітина ділиться на дві, і кожна клітина отримує по одному елементу з кожної пари хромосом.

Незважаючи на те, що в усіх клітинах організму присутній повний набір генів, лише мала частина геному, зазвичай, використовується або експресується будь-якою окремою клітиною. Наприклад, клітина печінки експресує набір генів, який відрізняється від того, який експресується клітиною шкіри.

Після того, як послідовність розшифрована, з'ясовані усі нуклеотиди та їх порядок, необхідно із цих даних отримати значущу інформацію. Існують комп'ютерні програми, за допомогою яких можна визначити ділянки, що кодують білки. У послідовності можна виділити такі ділянки, що відповідають послідовностям уже відомих білків, та спрогнозувати їх функції. Але, нажаль, за даними, отриманими порівняльними методами, не завжди можливо визначити коли, за яких обставин, у яких тканинах і окремих клітинах організму цей білок утворюється. Як дія певного білка вписується у діяльність інших білків, можливо, змінюючи її? Що станеться, якщо цього білка не виявиться у потрібному місці у потрібний час? На усі ці питання можна відповісти за допомогою лабораторних дослідів.

Важко переоцінити значимість розшифровки геному, вона надала величезне поле діяльності для біоінформатики. За допомогою методів біоінформатики стало можливим порівняти окремі ділянки хромосом між собою або з геномом інших тварин, результати цих порівнянь незамінні при вивченні еволюції та інших тривалих у часі процесів, при узагальненні і пошуку подібних послідовностей. Біоінформатика дозволяє зберегти багато сил і часу науковців, наприклад, знаючи функцію білка не обов'язково розшифровувати його послідовність цілком, оскільки кожен білок унікальний, за допомогою частини

його послідовності можна легко знайти повну послідовності білка і його «родичів» за допомогою біоінформаційних баз даних.

2. 5 Синтез білка

Розглянемо детальніше процес, за допомогою якого генетична інформація, зашифрована в молекулі ДНК реалізується в білках. У організмі еукаріот, такий процес відбувається у декілька етапів:

1. *Транскрипція* – це зчитування генетичної інформації, зашифрованої у молекулах ДНК, і запис цієї інформації у молекули іРНК. Клітинний механізм розпізнає початок окремого гена або цілої групи генів завдяки промотору. Промотор — це ділянка ДНК, розташована перед кожним геном, яка вказує на те місце в гені, з якого треба починати синтез РНК. Кодон AUG, який кодує метіонин, сигналізує про початок кодуючої області РНК. Основний фермент транскрипції РНК-полімераза приєднується до промотору. По мірі просування РНК-полімерази по кодуючому ланцюгу ДНК, рибонуклеїди по принципу компліментарності приєднуються до ланцюга ДНК, в результаті утворюється незріла іРНК, що містить як кодуючи, так і не кодуючи нуклеотидні ділянки. У результаті утвориться іРНК, що має точно таку ж послідовність нуклеотидів, як одна з ниток ділянки ДНК, що містить ген, який копіюється (але із заміною основи Т на U). Ця нитка комплементарна другій нитці ДНК.

2. *Процесинг* – дозрівання молекули РНК. Гени еукаріотів мають велику довжину і складну будову. Як уже зазначалося раніше, вони включають у себе окрім кодуючих послідовностей – екзонів, багаточисельні ділянки – інтрони. Продукти транскрипції модифікуються перед тим, як з них утворюються зрілі матричні РНК (мРНК, синонім — інформаційна РНК або іРНК). Процесинг включає у себе модифікацію 5-го та 3-го кінців і сплайсинг.

Модифікація 5-го кінця відбувається після полімеризації її перших 20-30 нуклеотидів. До цього кінця приєднується гуанозиндифосфат, що складається з пірофосфату, цукру пентози і рибози та азотистої основи гуаніну. Потім гуанін

метилується. Таке приєднання забезпечує ініціацію трансляції, оскільки лише за цієї умови рибосома розпізнає кодони-ініціатори AUG та GUG, захист мРНК від нуклеаз клітини, тим самим подовжуючи час її життя, роботу ферментативної системи, що проводить сплайсинг. У більшості транскриптів 3-й кінець також модифікується. За допомогою ферменту нарощується «хвіст», що складається з 100-200 залишків аденілової кислоти, що полегшує вихід мРНК з ядра і уповільнює її гідроліз у цитоплазмі.

Важлива частина процесингу – *сплайсинг*. У процесі сплайсингу приймають участь малі ядерні РНК (мяРНК), які формують сплайсому. ЇЇ каталітична активність обумовлена РНК-складовими (такі РНК називаються рибозимами). На кінцях інтронів знаходяться специфічні послідовності нуклеотидів – *сайти сплайсингу*, що забезпечують видалення інтронів. Формування сплайсоми відбувається шляхом з'єднання мяРНК та сайтів сплайсингу. При цьому кінці екзонів зближуються і з'єднуються, а інтрони видаляються.

Транскрипція і процесінг відбуваються у ядрі клітини. Потім зріла іРНК, яка складається тільки з екзонів, крізь пори у мембрані ядра виходить у цитоплазму, а потім починається трансляція.

3. *Трансляція* – процес синтезу поліпептидного ланцюга у відповідності з інформацією, яка закодована у рибосомній РНК.

Процес синтезу білка відбувається в клітинних структурах – рибосомах. Рибосоми побудовані з білків, прикріплених до каркаса з рибосомної РНК, або рРНК. На «вхід» рибосоми подається молекула іРНК, а також молекули тРНК, а з «виходу» знімається готовий поліпептидний ланцюг.

Молекули тРНК забезпечують зв'язок між кодоном і певною амінокислотою, яка ним кодується. На одній стороні кожної молекули тРНК перебуває структура, що має високий ступінь спорідненості з певним кодоном, а на іншій стороні — структура, що легко зв'язується з відповідною амінокислотою. Коли іРНК просувається через внутрішню частину рибосоми, тРНК розпізнає поточний кодон — кодон у іРНК, що перебуває тепер усередині рибосоми, — комплементарно зв'язується з ним, приносячи з собою відповідну амінокислоту

(«поблизу» активної рибосоми завжди є великий запас амінокислот). У цей момент тривимірне положення всіх цих молекул таке, що, як тільки тРНК зв'язується зі своїм кодоном, прикріплена до неї амінокислота виявляється в безпосередній близькості з попередньою амінокислотою у формованому поліпептидному ланцюзі. Тоді спеціальний фермент каталізує приєднання нової амінокислоти до білкового ланцюга й звільняє її від тРНК. У ході даного процесу поліпептидний ланцюг послідовно нарощується залишок за залишком. Коли з'являється стоп-кодон, йому не відповідає ніяка тРНК і синтез припиняється. ІРНК звільняється й розщеплюється клітинними механізмами до рибонуклеотидів, які у свій час будуть повторно використані для синтезу нової молекули РНК.

На перший погляд може здатися, що існує стільки ж видів молекул тРНК, скільки кодонів, але це не вірно. Фактичне число молекул тРНК варіює в різних видів організмів. Наприклад, мітохондрія має 22 вида тРНК, а бактерія *E. coli* – приблизно 40 видів. Деякі кодони не представлені молекулами тРНК, а деякі види тРНК можуть зв'язувати більше одного кодона.

Запитання до розділу 2

1. Яку будову і склад має молекула ДНК?
2. За допомогою яких зв’язків нуклеотиди з’єднані у ДНК?
3. Сформулюйте принцип компліментарності ланцюгів ДНК.
4. Запишіть ланцюг комплементарний до ланцюга TCCCGAAGGTTCTAG.
5. Як називається і в чому полягає процес точного самовідтворення ДНК?
6. Що спільного і відмінного мають молекули ДНК і РНК?
7. Які види РНК Вам відомі, які функції вони виконують?
8. Опишіть будову і функції білків.
9. Які джерела надають інформацію про первинну та просторову структуру білка?
10. Які методи дослідження білків Вам відомі?

11. Що називається геном?
12. Що називається кодоном і генетичним кодом? Назвіть характерні особливості генетичного коду.
13. Запишіть послідовність залишків амінокислот білка, що відповідає наступній послідовності РНК (зчитування починати з першого символа): UCCCGAAGGUUCUUAUAA.
14. Що називається геномом?
15. Що таке транскрипція і як відбувається цей процес?
16. Яку роль відіграє сплайсинг у прцесі синтезу білка?

Література до розділу 2

1. Ю.А. Овчинников, Биоорганическая химия, Москва, Просвещение, 1987 г., 815 с.
2. Шабарова З.А., Богданов А.А. Химия нуклеиновых кислот и их компанентов. – М.: Химия, 1978.
3. А. Ленинджер, Основы биохимии, Т.1 [пер. с англ. В.В. Борисова, М.Д. Гроздовой, С.Н. Преображенский, под ред. В.А Энгельгардта, Я.М. Варшавского], Москва, мир, 1985, 367 с.
4. С. Игнасимуту, Основы биоинформатики, [пер. с англ. А.А. Чумичкина], М.-Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. – 320 с.
5. Сетубал Ж., Мейданис Ж. Введение в вычислительную и молекулярную биологию. – М. – Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. – 420 с.
6. Глазко В.И., Глазко Г.В. Введение в генетику, биоинформатика, ДНК-технология, генная терапия, ДНК-экология, протеомика, метаболика – К.:КВІЦ, 2003.

7. Lesk M. Introduction to Bioinformatics – Oxford University Press Inc. New York, 2002.
8. Албертс Б. Молекулярная биология клетки: в 3 т. – М.: Мир, 1994.

РОЗДІЛ 3

БІОІНФОРМАЦІЙНІ БАЗИ ДАНИХ

Основою успіху біоінформатики є не тільки великий інтерес до неї науковців та промисловців, але й використання нею нетрадиційних для математики, фізики або хімії підходів. По-перше, це створення і широке використання загальнодоступних молекулярно-біологічних БД, які є одним з основних її інструментів. По-друге, це надання певним емпіричним правилам, що одержані шляхом статистичного аналізу інформації з відповідних БД, статусу «законів», наприклад, «золоті правила», які дозволяють робити висновки про фізико-хімічні та функціональні характеристики складних молекул на основі їх подібності до інших молекул з уже відомими властивостями.

3.1 Призначення та класифікація баз даних

З 1982 р. по молекулярній біології, зокрема, нуклеотидним та амінокислотним послідовностям, створено декілька тисяч БД. Серед них є електронні банки даних як загального призначення, так і вузькоспеціалізовані, а також БД наукових робіт з біології та медицини (Medline/PubMed та інші). Взагалі, **база даних** – це сукупність пов’язаної інформації, що об’єднана за певними ознаками. Більшість БД для збереження даних використовують таблиці. Кожна таблиця складається з рядків та стовпчиків, які називаються записами та полями, відповідно. Один запис може містити багато однакових полів, в які заноситься різна інформація.

Біоінформаційні БД забезпечують зручне та ефективне зберігання великої кількості інформації, систематизацію амінокислотних і нуклеотидних послідовностей, спрямовану на їх порівняльний аналіз, зокрема для:

- транслювання амінокислотних послідовностей білків;

- ідентифікації організмів, їх таксономічної приналежності і рівнів еволюційного розвитку, побудови філогенетичних дерев;
- задач генної інженерії;
- виявлення у неперервній послідовності символів окремих структурних одиниць та визначення їх функціонального навантаження;
- розшифровки просторової структури білків;
- виявлення структурно-функціональних взаємозв'язків груп білків;
- виявлення генів, які кодують макромолекули – потенційні мішені дії нових ліків та їх синтез (drug-desing).

Основним призначенням БД, крім збереження інформації, є швидкий пошук та цілеспрямоване структурування останньої. Електронні БД також повинні забезпечувати користувача засобами для управління всіма їх даними та інструментами для аналізу відповідної інформації.

Існує багато різних типів баз даних, які відрізняються за джерелом надходження інформації (архівні БД, БД, що куруються, похідні БД та інші) або за тематикою (нуклеотидні послідовності або цілі геноми, амінокислотні послідовності або просторова структура білків, конкретні організми або наукова література та інше).

Архівні, або первинні БД містять необроблені дані у тій формі, у якій вони були отримані із джерела. Існують також і вторинні БД (БД, що куруються), які містять відібрану після аналізу з архівних БД інформацію. Так БД GenBank є найбільшою архівною БД нуклеотидних послідовностей, а БД Swissprot, що курується, є найбільш достовірною БД по білках.

Ще в один тип БД можна виділити інтегровані БД, у яких зібрана уся інформація з певної тематики як з БД, що куруються, так і з БД, які не куруються. Якщо у таку БД ввести назву гену можна знайти усе, що про нього відомо – у яких організмах він зустрічається, у якому місці він локалізований, які функції виконує, яку просторову структуру він має і таке

інше. Такою БД є NCBI Entrez – доступ до інформації про нуклеотидні і амінокислотні послідовності і їх структурах.

За тематикою БД, як уже зазначалося, часто поділяють на два типи, а саме на бази даних загального та спеціального призначення. БД ДНК, білків, вуглеводів та інші є базами загального призначення. У якості прикладів спеціалізованих БД можна навести БД поліморфізмов окремих нуклеотидів, послідовностей, що характеризують геном, резус даних про імуногенні білки і таке інше. БД загального призначення, у свою чергу, часто поділяють на бази даних послідовностей і бази даних структур. Бази даних структур містять записи окремих послідовностей – нуклеотидів, амінокислот, білків. Бази даних структур містять записи окремих послідовностей, структури макромолекул яких визначені біохімічними методами (наприклад, БД Protein 3D structure).

Записи БД містять нові експериментальні результати і додаткові відомості у формі анотацій. Анотації дають інформацію про джерела даних і методи отримання цих даних. Також надається інформація про дослідників і перелік публікацій по даному питанню. Не менш важливим є те, що записи забезпечують посилання на відповідні записи інших БД.

Кожна база даних має свій формат представлення інформації, крім того, деякі з них виникли досить давно та зараз рідко використовуються. Для переводу з формату в формат існують різні програми-конвертори. Крім форматів EMBL, GeneBank, Swiss-Prot та інших БД, орієнтованих на сприйняття інформації людиною, необхідно знати один з найбільш поширених комп’ютерно-орієнтованих форматів – FASTA-формат, який слугує для представлення нуклеотидних та амінокислотних послідовностей у вигляді, зручному для обробки на обчислювальних машинах. Важливо також вміти переводити в нього данні з більш високорівневих форматів.

3.2 Методики пошуку інформації у БД

Усі існуючі бази даних надають можливісь роботи з ними через Internet та практично усі вони використовують стандартні методики пошуку, наприклад, можливість роботи з пошуковими системами:

Entrez (пошук по назві, номеру, організму, автору і т. ін.). Забезпечує доступ до амінокислотних і нуклеотидних послідовностей, їх тривимірних структур, а також до повних секвенованих геномів, надає графічне відображення генів. Практично для кожної послідовності можна підібрати подібні послідовності та вже розраховані і визначені дво- та тривимірні структури, що відносяться до даної послідовності.

BLAST (basic local alignment search tool, пошук за подібністю) – порівнює надану інформацію з послідовностями, що вже є в базі для пошуку подібних послідовностей. Є різні модифікації програми BLAST: BLASTp (вирівнювання амінокислотних послідовностей), BLASTn (вирівнювання нуклеотидних послідовностей), BLASTx (вирівнювання всіх можливих транслятів нашої нуклеотидної послідовності проти банка амінокислотних послідовностей), TBLASTx (вирівнювання всіх можливих транслятів нашої нуклеотидної послідовності проти всіх транслятів банка нуклеотидних послідовностей).

Якщо при пошуку за допомогою BLAST було віднайдено декілька подібних до вихідної послідовностей (для кожної з яких побудовано тільки парне вирівнювання з досліджуваною послідовністю), то виникає задача написати всі ці послідовності одна під одною, щоб визначити, в якій мірі вони співпадають, що в них консервативно (стійко) повторюється, а що ні. Ця задача називається множинним вирівнюванням.

Offline-інтерфейс – спочатку з мережі Інтернет на локальний комп’ютер скачується частина бази даних, потім з цією частиною проводиться подальша робота.

Режим клієнт-сервер – на локальному комп'ютері встановлюється програма математичної обробки нуклеотидних послідовностей або послідовностей амінокислот, далі дана програма з'єднується з сервером бази даних і обробляє інформацію без скачування останньої на локальний комп'ютер. В той же час інтенсивно розвиваються системи обробки інформації та пошукові системи, що збирають і обробляють інформацію відповідно до запитів користувачів.

Програмне забезпечення баз даних повинно задовільняти наступним функціональним вимогам:

- Об'єм баз даних повинний бути практично не обмеженим (тобто обмежений лише параметрами апаратних засобів).
- БД повинна бути достатньо гнучкою для забезпечення проходження процесу перебудови по мірі її заповнення, так як попереднє проектування детальної структури бази даних є неможливим.
- БД повинні бути інтегровані з іншими БД та підтримувати не лише стандартні мультимедійні формати, але й ряд спеціальних гіпермедіа-середовищ (просторові структури молекул, хімічні структурні формули та ін.).
- Експлуатація та поповнення баз даних через комп'ютерні мережі має бути легко доступним та зрозумілим для користувачів, які не мають комп'ютерної підготовки (біологи, медики).

3.3. Характеристики біоінформаційних ресурсів

Однією з найвідоміших міжнародних організацій з тих, що створюють інструменти для аналізу інформації та здійснюють нагляд за наповненням баз даних біоінформаційного напряму, є Міжнародна система баз даних нуклеотидних послідовностей (**International Nucleotide Sequence Database Collaboration (INSDC)** –http://www.ncbi.nlm.nih.gov/collab), яка об'єднує три установи, нуклеотидні БД яких мають різний набір сервісів:

NCBI – National Center for Biotechnology Information, USA (БД GenBank містить більше 8×10^7 записів), http://www.ncbi.nlm.nih.gov/ ,

EBI – European Bioinformatics Institute, United Kingdom (БД EMBL містить біля 8×10^7 нуклеотидних послідовностей, а TrEMBL – більше 5×10^6 автоматично трансльованих з даних EMBL амінокислотних послідовностей), http://www.ebi.ac.uk/ ,

NIG – National Institute of Genetics, Japan (БД DDBJ містить біля 8×10^7 записів), http://www.ddbj.nig.ac.jp/ .

До найбільш відомих організацій також відносяться RCSB – Research Collaboratory for Structural Bioinformatics, USA (БД PDB – Protein data bank – біля 5×10^4 просторових структур білків), http://www.rcsb.org/, та SIB – Swiss Institute of Bioinformatics, Switzerland (БД Swissprot містить більше 3×10^5 записів білків), http://www.isb-sib.ch/. Дані вказані за станом на березень 2008 року.

Розглянемо більш докладно деякі з БД:

– *БД нуклеотидних послідовностей:*

EMBL (European Molecular Biology Laboratory, Geyzelberg, Germany, http://www.ebi.ac.uk/embl/) – заснована у 1982 р. БД нуклеотидних послідовностей. Поповнюється безпосередньо авторами, що визначили первинну структуру фрагмента ДНК чи РНК. В форматі БД EMBL представлена інформація про нуклеотидну послідовність, її функціональну розмітку, посилання на відповідні експериментальні дані та статті. Картка EMBL представляє з себе текстовий файл з обмеженою довжиною строки. Ключові слова, що визначають характер представленої в різних областях картки інформації є жорстко заданими. Для картки EMBL ключові слова створюють двосимвольні послідовності на початку строки та служать для розширення можливостей двовимірної БД. Список ключових слів наведений в таблиці 3.1.

Ключове слово	**Description**	**Опис**
Identification	Contains identifying information and characteristics of the sequence	Унікальний ідентифікатор послідовності, містить інформацію про послідовність, може змінюватися від одного випуску БД до іншого
Accession number(s)	Release-to-release stable identifiers	Унікальний ідентифікатор послідовності, що не змінюється від випуску до випуску БД
Date	When the entry was created, or when the sequence or annotation was modified	Дата створення запису та останньої її модифікації
Description	The name of the protein, often a function indicator	Назва білка, що часто вказує на його функцію
Gene name(s)	The gene(s) that code for the protein	Ім’я гена (-ів), що кодують білок
Organism species	The organism from which the sequence is derived	Організм, в якому знайдений білок
Organelle	If the sequence is non-chromosomal in origin	Якщо ген білка не знаходиться в хромосомі
Organism classification	The taxonomic class to which the organism belongs	Таксономічна класифікація організму
Taxonomy cross-reference(s)	The NCBI TaxID for the OC line	Ідентифікатор таксону
Reference number	The sequential number of the literature citation within the entry	Номер посилання на наукову публікацію по даному білку
Reference position	The type of data, and the position in the sequence to which the citation refers	Тип інформації в даному посиланні та фрагмент послідовності, до якого вона відноситься.
Reference comment(s)	Comments relevant to the reference cited	Коментарі до роботи, що цитується
Reference cross-reference(s)	Bibliographic cross-reference, such as PubMed ID	Бібліографічні перехресні посилання, наприклад, ідентифікатор статті в БД PubMed
Reference	Authors of the citation	Автори публікації

authors		
Reference title	Title of the citation	Назва статті
Reference location	Source of the citation, such as journal, book, or unpublished data	Джерело цитування: журнал, книга чи неопубліковані данні
Comments or notes	Free text notes about the protein	Будь-які інші коментарі по даному білку
Database cross-references	Pointers to sources or related information for the entry	Гіперпосилання на відповідні записи інших БД
Keywords	Indexable indicator of function, structure, or other information	Ключові слова (індексований укажчик функції, структури та іншої інформації по білку)
Feature table data	Annotation of specific residues of the sequence	Опис особливих позицій чи властивостей послідовності, що вказує тип властивості (Key name)
Sequence header	Marks the beginning of the sequence and provides summary data	Мітка початку послідовності, де також знаходяться деякі її загальні характеристики
Sequence data	The sequence itself	Амінокислотна послідовність
Termination line	//	Мітка кінця запису

Таблиця 3.1. Ключові слова БД EMBL та опис відповідних полів.

GenBank (National Centre for Biotechnology Information, Los–Alamos, USA, http://www.ncbi.nlm.nih.gov/) – заснована у 1982 р. БД генетичних послідовностей, містить анотовану колекцію всіх загальнодоступних послідовностей ДНК, РНК та білків разом з літературними посиланнями.

Поповнюється раз на 2 місяці. Станом на жовтень 1999 р. містить 3,841,163,011 нуклеотидних пар в 4,864,570 записах послідовностей, на вересень 2002 р. – 20,017,246,707 нуклеотидних пар в 8,293,265 записах послідовностей, на лютий 2004 р. – 37,893,844,733 нуклеотидних пар в 32,549,400 записах послідовностей, на лютий 2008 р. - 85,759,586,764 нуклеотидних пар в 82,853,685 записах послідовностей.

DDBJ (Center for Information Biology (cib), Dna Data Bank of Japan, http://www.ddbj.nig.ac.jp/) – заснована у 1984 р. БД генетичних послідовностей. Збір інформації проводиться в першу чергу у японських вчених та з літератури. 75% зібраних послідовностей являють собою частково секвеновані фрагменти ДНК з декількох сотень експресованих генів, так званих EST (Expressed Sequence Tags).

Генэкспресс (ВИНИТИ–ИМГ Россия, ftp://ftp.infobiogen.fr/pub/db/Genexpress) – БД генетичних послідовностей. Містить генетичні послідовності по 10,000 генів людини. Оновлюється раз на два місяці.

– *БД білкових послідовностей та 3d структур:*

Swissprot (Department of Medical Biochemistry of the University, Switseland, Geneva, http://www.expasy.ch/sprot/sprot-top.html) – БД, що містить анотовані амінокислотні послідовності, трансльовані з нуклеотидних послідовностей EMBL; адаптовані послідовності з PIR; а також послідовності опубліковані в літературі і надіслані безпосередньо авторами. Містить високоякісні анотації без збиткової інформації, посилання на споріднені бази даних (EMBL, Prosite, PDB). Кожна анотація містить опис функції білка, його доменної структури, особливостей пост-трансляційної модифікації. Оновлюється щотижня. Для академічних користувачів є безкоштовною.

PIR (protein information resource, National Biomedical Research Foundation, USA, http://www-nbrf.georgetown. edu/pirwww/dbinfo/pirpsd.html) – БД, що містить інформацію щодо білків для яких відомі нуклеотидні послідовності. Пошук організовано як по таксономії так і гомології. Має низький рівень зайвої інформації. Поповнюється щотижня.

MMDB (Molecular Modelling Database, Georgetown University, USA, http://www.ncbi.nlm.nih.gov/Structure/) – БД, що містить просторові структури білків, визначені дослідним шляхом (рентгеноструктурною кристало-графією та ЯМР-спектроскопією), надає інформацію про біологічну функцію та

механізми, що з нею пов'язані; еволюційну історію і взаємозв'язок між макромолекулами. Входить до складу PDB, що містить також теоретичні моделі. Всі структури цієї бази даних мають первинні структури в NCBI. Оновлюється щоденно.

ENZYME (http://www.expasy.ch/enzyme**) –** БД, що містить інформацію щодо номенклатури ферментів, і описує всі типи білків, яким присвоєно номер ЕС (Enzyme Commission). Пошук реалізовано по ЕС-номеру, класам ферментів, хімічним компонентам, по кофакторам, по назвам хвороб, пов'язаних з ферментом.

– *Спеціалізовані біоінформаційні ресурси:*

OMIM – (John Hopkins University, USA, http://www3.ncbi.nlm.nih.gov/omim) – БД по генах людини та пов'язаними з ними хворобами. Підтримується NCBI. Оновлюється щоденно.

PMB (Protein Mutant Database, Japan, http://www.genome.ad.jp/htbin/www_bfind?pmd) – БД призначена для досліджень в області білкової інженерії. Одиницею інформації є не білок, а одна стаття, присвячена мутації.

Genes Kegg (Kioto, Japan, http://www.genome.ad.jp/kegg/) – БД по систематичному аналізу функцій генів Складається із шести баз даних – метаболічних шляхів (PATHWAY), генів (GENES), лігандів (LIGAND), дослідних даних по експресії генів (EXPRESSION и BRITE), білків (SSDB). Забезпечує можливість роботи з усіма крупними світовими інформаційними ресурсами. Оновлюється щоденно.

Оволодіння навичками користування інтернет-ресурсами молекулярної біології відкриває широкі можливості у використанні біоінформатики (або обчислювальної молекулярної біології) не лише для пошуку і аналізу вже існуючої інформації, але й для отримання нових знань з меншими затратами

матеріальних та часових ресурсів порівняно з фізико-хімічними дослідженнями.

Таким чином, окрім високого професійного рівня в області біології, ключовим моментом в оволодінні всіма можливостями сучасних біоінформаційних технологій є наявність у фахівця навичок швидкого та ефективного пошуку інформації в спеціалізованих БД через мережу Інтернет. В цьому аспекті слід згадати журнал Nucleic Acid Research (NAR), який на протязі останніх кількох років перший номер року присвячував опису біоінформаційних БД, а з 2007 року замість цього випускає спеціальний додатковий номер.

Запитання до розділу 3

1. Дати визначення бази даних (БД).
2. Які функції виконують БД ?
3. Яким чином, зазвичай, класифікують БД?
4. На які типи за тематикою поділяються БД?
5. Наведіть приклади архівних БД, та БД, що куруються.
6. Які методики пошуку інформації у БД Вам відомі?
7. Розкрийте суть роботи програми BLAST.
8. Перерахуйте найбільш відомі біоінформаційні ресурси.
9. Охарактеризуйте основні характеристики БД нуклеотидних послідовностей EMBL.
10. Що Вам відомо про БД білкових послідовностей та 3d структур?
11. Які біоінформаційні ресурси можна віднести до спеціалізованих?
12. Яким основним вимогам повинно задовільняти програмне забезпечення баз даних.

Література до розділу 3

1. 1Дурбин Р., Эдди Ш., Крог А., Митчисон Г. Анализ биологических последовательностей. – М. – Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2006. – 480 с.
2. Сетубал Ж., Мейданис Ж. Введение в вычислительную и молекулярную биологию. – М. – Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. – 420 с.
3. Глазко В.И., Глазко Г.В. Введение в генетику, биоинформатика, ДНК-технология, генная терапия, ДНК-экология, протеомика, метаболика – К.:КВІЦ, 2003.
4. Bioinformatics: Sequence, structure and database – Oxford University Press, 2001.

РОЗДІЛ 4

МЕТОДИ ВИРІВНЮВАННЯ НУКЛЕОТИДНИХ І АМІНОКИСЛОТНИХ ПОСЛІДОВНОСТЕЙ

Дослідження нових нуклеотидних послідовностей ДНК і РНК, а також амінокислотних послідовностей білків відбувається на основі уже відомих, розшифрованих послідовностей. Цей факт є вихідним для будь-якого аналізу послідовностей. Це означає, що у разі подібності недавно відкритих послідовностей з послідовностями молекул, структура або функція яких уже відома, ці властивості з деякою мірою точності можна приписати новій послідовності також. Таким чином, порівняння послідовностей необхідне для виявлення загальних структурних і функціональних особливостей та для дослідження еволюційних зв'язків.

4.1 Глобальне вирівнювання

Першим кроком аналізу послідовностей є визначення подібності послідовностей. Якщо дві послідовності дуже довгі, важко визначити, чи насправді вони подібні. Для того, щоб дізнатися про їх подібність необхідно певним чином їх вирівняти. *Вирівнювання послідовностей* – це схема запису однієї послідовності над іншою, таким чином, щоб їх елементи у певних позиціях максимально відповідали спільному походженню. Якщо послідовності походять від спільного пращура, їх залишки можуть замінюватися (один залишок замінено на інший), окрім цього у процесі еволюції у послідовностях можуть накопичитися зміни наступних типів: вставки (у послідовності з'явилися нові додаткові залишки) і випадіння (деякі залишки зникають). Враховуючи зазначене, у послідовностях можуть міститися не лише зайві символи, але і пробіли.

Розглянемо дві нуклеотидні послідовності: AGTCTAGTCA і AGCTAGACA. Вони досить схожі, що стає очевидним, коли ми записуємо послідовності одну над іншою:

AGTCTAGTCA

AG_CTAGACA.

Розходження пов'язані із зайвим символом T у третій позиції та заміною символу T на символ A у восьмій позиції зліва направо. Помітивши це, ми ввели пробіл в другій послідовності, щоб вирівняти співпадаючі літери ліворуч і праворуч від пробілу.

Тепер розглянемо дві інші нуклеотидні послідовності :

AGCAATCGCTG

AGAAATCG

Ці послідовності мають різні довжини. Ми визначимо вирівнювання як розташування пробілів у довільних місцях одної, або обох послідовностей так, щоб можна було знайти найбільш співпадаючі ділянки, і так, щоб вони мали однакову довжину. Послідовності однакової довжини ми можемо записати тепер одну над іншою так, щоб установити відповідність між пробілами – символами однієї послідовності й пробілами – символами іншої. Спробуємо вирівняти послідовності наступним чином:

AGCAATCGCTG

AG_ AAAT_C _G

Дане вирівнювання може означати, що наприклад, фрагмент AGCAATC у процесі еволюції міг зазнати випадіння 2-го символу C, та заміни 6-го та 7-го символів зліва направо. Але у другій послідовності можна розставити пробіли і іншим чином, так, що таке вирівнювання також здавалось би припустимим:

AGCAATCGCTG

AGAAATC_ _ _G

Таке вирівнювання може означати, що фрагмент AGCAATC у процесі еволюції міг зазнати заміни 3-го символу C на символ A зліва направо. Яким же чином можна визначити яке вирівнювання краще? Відповідь залежить від

поставлених задач. У загальному випадку, кращим, або «оптимальним» вирівнюванням називається те, яке має більшу вагу. Стовпчик вирівнювання отримує, залежно від змісту, певний бал, і загальна вага вирівнювання буде сумою балів по усіх стовпчиках. Якщо два символи в стовпчику збігаються, то бал дорівнює 1 (збіг). У випадку розбіжності бал дорівнює –1, а пробіл у стовпчику означає бал – 2. Найкращим вирівнюванням буде те, що максимізує загальну вагу вирівнювання. Цю максимальну вагу будемо називати подібністю двох послідовностей *s*, *t* і позначимо $\omega = a(m,n)$, де *n* та *m* – кількість елементів в послідовностях *s* та *t*. У загальному випадку, може бути декілька вирівнювань з максимальною вагою.

Тепер, щоб дати відповідь на запитання попереднього прикладу розрахуємо вагу обох вирівнювань. У першому випадку шість стовпчиків з однаковими символами, два – з різними і три пробіли. Тоді сумарна вага буде дорівнювати $\omega = 6 \cdot 1 + 2 \cdot (-1) + 3 \cdot (-2) = -2$. У другому випадку сім стовпчиків з однаковими символами, один – з різними і три пробіли. Тоді сумарна вага буде дорівнювати $\omega = 7 \cdot 1 + 1 \cdot (-1) + 3 \cdot (-2) = 0$. Легко бачити, що друге вирівнювання має більшу сумарну вагу, а отже є оптимальним.

Величини –1, –2, 1 були обрані тому, що вони досить часто використовуються на практиці, коли штрафуються пробіли й розбіжності. Далі ми будемо розглядати і інші системи штрафів.

Таким чином, природний підхід для з'ясування подібності між двома послідовностями полягає у побудові всіх можливих вирівнювань та виборі найкращого – тобто того, що має максимальну вагу.

Існує теорема, доведена Лак'є: нехай $f(n,m)$ - кількість всіх можливих вирівнювань послідовностей довжин *n* і *m*. Тоді для досить великих *n* справедливе співвідношення

$$f(n,m) \cong \left(1+\sqrt{2}\right)^{2n+1}\sqrt{n}$$

Тобто для двох послідовностей, довжиною 1000 існує

$$f(1000,1000) \cong \left(1+\sqrt{2}\right)^{2001}\sqrt{1000} = 10^{767,4...}$$

можливих вирівнювань, для прикладу, кількість елементарних часток у Всесвіті порядку 10^{80}, а число Авогадро $6{,}02 \cdot 10^{23}$ моль$^{-1}$.

Таким чином, очевидно, що прямий перебір всіх вирівнювань неможливий. Тому необхідно використовувати оптимальний алгоритм, який би дозволив істотно зменшити час обчислень.

Розглянемо існуючі підходи до вирівнювання, засновані на техніці динамічного програмування. *Алгоритмами динамічного програмування* називаються алгоритми, робота яких заснована на оптимізації, тобто, зменшенні числа даних (у даному випадку використовуються підрядки). Основна ідея полягає в одержанні розв'язку "рекурсивним" способом, тобто розв'язок задачі будується на основі вже отриманого розв'язку для об'єкта меншого розміру. Для двох послідовностей s і t замість того, щоб шукати вирівнювання (s,t) цілком, ми шукаємо вирівнювання між всіма можливими префіксами (s,t). Префікс – це підпослідовність послідовності, яка починається з першого символу послідовності і закінчується на довільному її символі. Спочатку відбувається вирівнювання між короткими префіксами, а отримані результати використовуються для більш довгих.

Вирівнювання, засновані на оптимізації явної функції премій за збіги й штрафів за розбіжності, називаються вирівнюваннями, заснованими на подібності (вирівнювання, засновані на метричних відстанях мають менш загальний характер).

Розглянемо задачу вирівнювання послідовностей, заснованому на алгоритмі глобального вирівнювання.

Нехай n – довжина послідовності s і m – довжина послідовності t. Очевидно, що максимальна довжина, яка буде однаковою для обох вирівняних послідовностей дорівнює $n + m$.

Розглянемо алгоритм глобального вирівнювання на прикладі послідовностей s = CAAATC та t = CTC. Будемо застосовувати, згадану вище систему премій та штрафів:

	пробіл	A	T	G	C
пробіл	0	g =-2	-2	-2	-2
A	-2	1	-1	-1	-1
T	-2	-1	1	-1	-1
G	-2	-1	-1	1	-1
C	-2	-1	-1	-1	1

	пробіл	A	T	G	C
пробіл	0	g =0	0	0	0
A	0	20	10	5	5
T	0	10	20	5	5
G	0	5	5	20	10
C	0	5	5	10	20

	пробіл	A	T	G	C
пробіл	0	g =25	25	25	25
A	25	0	20	20	20
T	25	20	0	20	20
G	25	20	20	0	20
C	25	20	20	20	0

де штраф або премія за розташування в стовпчику вирівнювання знаків вказані на перетині відповідного рядка та стовпчика таблиці, наприклад *g=-2* штраф

стовпчику вирівнювання «символ-пробіл» або «пробіл-символ» в верхній таблиці. Наступна таблиця є прикладом матричної системи ваг, в якій враховується, що пурин-пуринові і пиримидин-пиримидинові мутації a⇔g і t⇔c є більш поширеними, ніж пурин-пиримидинові a⇔t, g⇔c. Перші дві таблиці відносяться до випадку, коли ставиться задача пошуку вирівнювання з максимальною вагою, так як найбільша позитивна премія надається за збіг символів в стовпчику вирівнювання, а від'ємними числами штрафуються блоки типу «символ-пробіл». Остання (третя) таблиця є прикладом системи ваг, в якій збіг символів не штрафується, а найбільший штраф надається за блоки типу «символ-пробіл». При цьому ставиться задача пошуку вирівнювання з найменшою вагою, тобто з найменшим значенням сумарного штрафу.

1) *Побудова таблиці премій і штрафів p*.

В першому горизонтальному рядку таблиці 4.1 розміщують послідовність *s*, в першому вертикальному стовпчику – послідовність *t*. У клітинках таблиці $p(i, j)$ записується премія або штраф в залежності від того, які символи розташовані у *i*-му рядку та *j*-му стовпчику таблиці. Якщо символи співпали, то на перетині відповідного рядка і стовпчика записується 1, у протилежному випадку –1.

i ↓	j →	*1*	*2*	*3*	*4*	*5*	*6*
	t \ *s*	**C**	**A**	**A**	**A**	**T**	**C**
1	**C**	1	–1	–1	–1	–1	1
2	**T**	–1	–1	–1	–1	1	–1
3	**C**	1	–1	–1	–1	–1	1

Таблиця 4.1. Масив *p* для послідовностей *s* = CAAATC та *t* = CTC.

2) *Ініціалізація масиву ваг вирівнювання префіксів a.*

Масив ваг вирівнювання префіксів – це двомірний масив розміром $(m+1)\times(n+1)$, рядки і стовпчики якого нумеруються, починаючи з 0 (таблиця 4.2). У масив *a* записуються ваги вирівнювання префіксів послідовностей *s* та *t*. На перетині *i*-го рядка та *j*-го стовпчика стоїть вага вирівнювання префіксів, довжинами *i* та *j*; $a(m,n)$ – вага вирівнювання в цілому.

Нульовий рядок і стовпчик ініціюються сумою штрафів за пробіли. Це відбувається тому, що якщо один рядок або стовпчик порожній, то можливе тільки одне вирівнювання: додається кількість пробілів, яке дорівнює довжині непорожнього рядка або стовпчика. У символьній формі можна записати:

$$\begin{array}{l} a(0,j)=g\cdot j, \\ a(i,0)=g\cdot i, \end{array} \quad \text{де} \quad \begin{array}{l} i=0,...m, \\ j=0,...n. \end{array}$$

3) *Розрахунок всіх елементів масиву a за рекурентною формулою.*

Розрахунок інших елементів масиву оснований на тому, що значення $a(i,j)$ можна обчислити, розглянувши всього три попередні значення: $a(i,j-1)$, $a(i-1,j-1)$ та $a(i,j-1)$. Це пов'язано з тим, що префікси довжинами *i* та *j* можна вирівняти тільки трьома способами, і кожний використовує одне з попередніх значень. Фактично, щоб одержати це вирівнювання, потрібно вибрати одну з наступних трьох можливостей:

а) вирівняти префікс послідовності *t*, довжиною *i*, та префікс послідовності *s*, довжиною $j-1$, і зіставити пробіл та *j*-й символ послідовності *s*;

t: *Вирівняний префікс, що мав довжину i* + *пробіл*

s: *Вирівняний префікс, що мав довжину* $j-1$ + *j-й символ*

б) вирівняти префікс послідовності *t*, довжиною $i-1$, та префікс послідовності *s*, довжиною $j-1$, і зіставити *i*-й символ послідовності *t* та *j*-й символ послідовності *s*;

t: *Вирівняний префікс, що мав довжину* $i-1$ + *i-й символ*

s: *Вирівняний префікс, що мав довжину* $j-1$ + *j-й символ*

в) вирівняти префікс послідовності *t*, довжиною $i-1$, та префікс послідовності *s*, довжиною *j,* і зіставити *i*-й символ послідовності *t* і пробіл;

t: *Вирівняний префікс, що мав довжину* $i-1$ + *i-й символ*

s: *Вирівняний префікс, що мав довжину* j + *пробіл*

Ці можливості вичерпують вибір, оскільки неможливо вирівнювати пробіли в останньому стовпчику вирівнювання. Ваги для вирівнювань префіксів меншого розміру вже містяться в масиві, якщо вибрано правильний порядок обчислень. Тому подібність може бути визначена наступною формулою:

$$a(i,j)=\max\begin{cases} a(i,j-1)+g, \\ a(i-1,j-1)+p(i,j), \\ a(i-1,j)+g. \end{cases} \qquad (4.2)$$

Масив можна заповнювати по рядках зліва направо, або по стовпчиках зверху вниз, це не має значення.

i ↓	*j* →	*0*	*1*	*2*	*3*	*4*	*5*	*6*
	t \ *s*	_	**C**	**A**	**A**	**A**	**T**	**C**
0	_	0	–2	–4	–6	–8	–10	–12
1	**C**	–2	1	–1	–3	–5	–7	–9
2	**T**	–4	–1	0	–2	–4	–4	–6
3	**C**	–6	–3	–2	–1	–3	–5	**–3**

Таблиця 4.2. Глобальне вирівнювання: масив ваг вирівнювання префіксів *a* для послідовностей *s* = CAAATC та *t* = CTC.

4) *Зворотній хід алгоритму*.

Оптимальна вага глобального вирівнювання послідовностей s та t в цілому записана у (m,n) клітинці масиву a, для нашого прикладу оптимальна вага $\omega = -3$. *Зворотній хід алгоритму глобального вирівнювання* полягає у тому, щоб, починаючи з елементу (m,n) і закінчуючи елементом $(0,0)$ визначити зворотній шлях, за допомогою якого і було отримано оптимальну вагу і, відповідно, до знайденого шляху розставити у послідовностях пробіли у потрібних місцях. Шлях, зазвичай, зображують за допомогою горизонтальних, вертикальних і діагональних стрілок, що сполучають елемент таблиці з тим елементом, за допомогою якого, його було розраховано по рекурентній формулі (Таблиця 4.3).

i ↓	j →	*0*	*1*	*2*	*3*	*4*	*5*	*6*
	$t \setminus s$	_	**C**	**A**	**A**	**A**	**T**	**C**
0	_	0	–2	–4	–6	–8	–10	–12
1	**C**	–2	↖ –1	← –1	← –3	← –5	–7	–9
2	**T**	–4	–1	0	–2	–4	↖ –4	–6
3	**C**	–6	–3	–2	–1	–3	–5	↖ –3

Таблиця 4.3. Глобальне вирівнювання: зворотній шлях оптимального вирівнювання.

Слідуючи за стрілками можна отримати вирівняні послідовності, якщо кожній стрілці співставити стовпчики вирівняних послідовностей:

↖ – символи у s та t послідовностях;

← – символ у вертикальній і пробіл у горизонтальній послідовності;

↑ – пробіл у вертикальній і символ у горизонтальній послідовності.

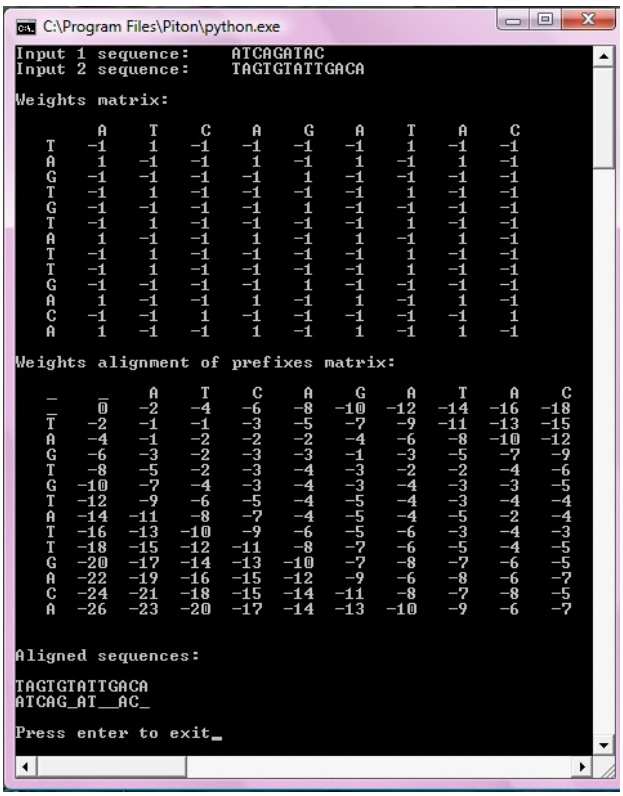
Таким чином для послідовностей $s = \text{CAAATC}$ та $t = \text{CTC}$ оптимальним вирівнюванням з вагою $\omega = -3$ буде наступне:

CAAATC
CT _ _ _C.

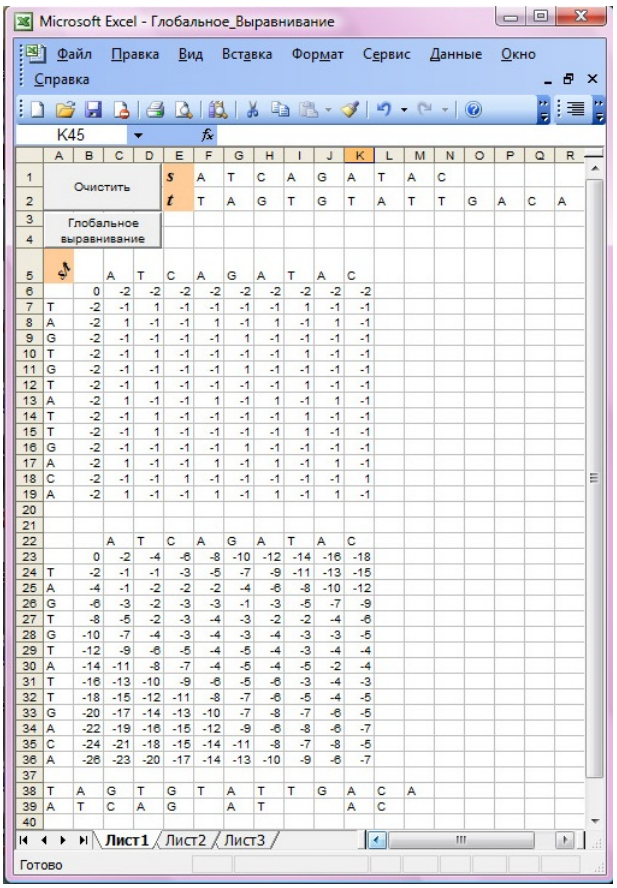
Очевидно, що для таких коротких модельних послідовностей, які ми розглядали для прикладу, не потрібно використовувати алгоритм глобального вирівнювання, тому що порівняти їх досить просто. У той же час для порівняння послідовностей на практиці алгоритм глобального вирівнювання незамінний. Варто зазначити, що для багатьох пар послідовностей може існувати декілька оптимальних вирівнювань, що мають однакову вагу, у таких випадках, для отримання необхідного результату потрібно враховувати особливості тої чи іншої задачі.

Глобальне вирівнювання білкових послідовностей дає дуже якісні результати, для послідовностей одного і того ж сімейства білків. Наприклад, білок цитохром С має майже однакову довжину в усіх організмах, які його виробляють. Тому можна очікувати відповідність цитохром з двох різних видів по всій довжині двох рядків. Те саме можна спостерігати для білків сімейства глобінів, таких як міоглобін і гемоглобін. При спробах визначити еволюцію цих білків за допомогою аналізу подібності і відмінності білкових послідовностей, зазвичай порівнюють білкові послідовності з одного сімейства білків, тому для вирішення даних задач застосування глобального вирівнювання достатньо виправдане і ефективне.

На рис. 4.1 приведені приклади реалізації алгоритму глобального вирівнювання на мові програмування: а) Python; б) Microsoft Visual Basic.



а)



б)

Рис. 4.1 Приклади реалізації алгоритму глобального вирівнювання на мові програмування: а) Python; б) Microsoft Visual Basic.

4.2 Локальне вирівнювання

З біологічної точки зору, однією з найбільш цікавих задач, пов’язаних із вирівнюваннями, є знаходження подібних ділянок заданих послідовностей *s* і *t*, вирівнювання яких мають найбільшу вагу. Для розв’язку такої задачі використовується алгоритм локального вирівнювання. Також локальне вирівнювання корисне, якщо послідовності дуже відрізняються по довжині, у такому випадку, коли немає потреби порівнювати довгу послідовність «від початку до кінця», значно інформативніше порівнювати лише її підрядки.

Локальне вирівнювання s, t – це вирівнювання підрядка *s* з підрядком *t*.

Алгоритм пошуку оптимального локального вирівнювання – це модифікація основного алгоритму для глобального вирівнювання. Як і раніше, вхідними даними є масив *a* розміром $(m+1)\times(n+1)$, тільки тепер кожне значення $a(i,j)$ дорівнює вазі вирівнювання між підрядками префіксу послідовності *t* довжиною *i*, та префіксу послідовності *s*, довжиною *j*.

Алгоритм локального вирівнювання

1) *Побудова таблиці премій і штрафів p.*

2) *Ініціалізація масиву ваг вирівнювання префіксів a.*

Локальне вирівнювання ставить за мету знайти у послідовностях найбільш подібні ділянки, вага вирівнювання яких має бути максимальною. Тому масив ваг *a* не може містити від'ємні значення. Нульовий рядок і стовпчик ініціюються нулями (таблиця 4.4). У символьній формі можна записати:

$$\begin{matrix} a(0,j)=0, \\ a(i,0)=0, \end{matrix} \quad \text{де} \quad \begin{matrix} i=0,...m, \\ j=0,...n. \end{matrix}$$

3) *Розрахунок всіх елементів масиву a за рекурентною формулою.*

$$a(i,j)=\max\begin{cases} 0, \\ a(i,j-1)+g, \\ a(i-1,j-1)+p(i,j), \\ a(i-1,j)+g. \end{cases} \tag{4.3}$$

Якщо вирівнювання у певний момент набуває від'ємної ваги, у такому випадку краще почати пошук нового вирівнювання підрядків, ніж продовжувати розпочате. Саме тому, у рекурентній формулі і з'явився четвертий варіант вибору 0, який виключає від'ємні значення елементів масиву *a* і означає кінець поточного локального вирівнювання.

i ↓	j→	*0*	*1*	*2*	*3*	*4*	*5*	*6*
	$t \setminus s$	_	**C**	**A**	**A**	**A**	**T**	**C**
0	_	0	0	0	0	0	0	0
1	**C**	0	↖1	0	0	0	0	↖1
2	**T**	0	0	0	0	0	↖1	0
3	**C**	0	↖1	0	0	0	0	↖2

Таблиця 4.4. Локальне вирівнювання: масив ваг вирівнювання префіксів *a* для послідовностей *s* = CAAATC та *t* = CTC.

4) *Зворотній хід алгоритму*.

Суттєва різниця між глобальним і локальним вирівнюваннями полягає у тому, що локальне вирівнювання підрядків може починатися не лише з нижньої правої клітинки, а і з будь-якого іншого місця у масиві. Що стосується кінця вирівнювання, згадаємо, що 0 означає кінець поточного локального вирівнювання і початок нового. Таким чином, масив *a* може містити велику кількість шляхів (їм відповідають різні вирівнювання підрядків з невід'ємною вагою), які розділені між собою нулями. З таблиці 4.4 видно, що для даного прикладу існує чотири варіанти вирівнювання, три з яких дають збіги :

C,

C.

Четвертому шляху відповідає комбінація:

TC,

TC.

Очевидно, що оптимальним вирівнюванням буде те, яке має більшу вагу, для даного приклада, це те, що має вагу 2, і відповідає найбільш подібною спільною ділянкою TC послідовностей *s* і *t*.

Вищезазначене означає, що *зворотній хід алгоритму локального вирівнювання* полягає у пошуку максимального елементу усього масиву *a* і

побудові зворотного шляху, рухаючись від максимального значення масиву *a* до першого нульового значення цього масиву (таблиця 4.5).

У багатьох біологічних задачах локальна подібність значно результативніше, аніж глобальна. Особливо важливу роль локальне вирівнювання відіграє при порівнянні білкових послідовностей, оскільки білки з дуже різних сімейств часто побудовані з одних і тих самих структурних і функціональних субодиниць. Прикладом такої ситуації можуть слугувати білки, закодовані гомеобоксними генами. Ці гени широко розповсюджені від дрозофіли до жаби і людини, вони регулюють ембріональний розвиток. Амінокислотні послідовності, які закодовані у цих генах дуже відрізняються у різних видів, окрім однієї ділянки, яка називається гомеодоменом. Гомеодомен складається приблизно з шістдесяти амінокислот, які утворюють частину регуляторного білка, який зв'язується з ДНК. Було встановлено, що при локальному вирівнюванні гомеодомени у комах і ссавців проявляють схожість від 50 до 95%. Зв'язування білка з ДНК відіграє провідну роль у регуляції ембріонального розвитку.

i ↓	*j*→	*0*	*1*	*2*	*3*	*4*	*5*	*6*
	t \ *s*	_	**C**	**A**	**A**	**A**	**T**	**C**
0	_	0	0	0	0	0	0	0
1	**C**	0	1	0	0	0	0	1
2	**T**	0	0	0	0	0	↖1	0
3	**C**	0	1	0	0	0	0	↖***2***

Таблиця 4.5. Локальне вирівнювання: зворотній шлях вирівнювання.

Таким чином, послідовність амінокислот у найбільш біологічно важливій частині цих білків висококонсерватина, у той час, як подібність інших ділянок незначна. У таких випадках, як розглянутий, локальне вирівнювання дає кращий спосіб порівняння амінокислотних послідовностей, ніж глобальне.

4.3 Псевдоглобльне вирівнювання

Псевдоглобальне вирівнювання – це таке вирівнювання, яке не враховує деякі кінцеві пробіли в послідовностях, тобто ті, які стоять перед першим або після останнього символу. Наприклад, всі пробіли у другій послідовності кінцеві, на відміну від єдиного пробілу в першій.

AGAC_CAGATTTCTGCCAG
AGACTCAG_ _ _ _ _ _ _ _ _ _

Відмітимо, що дві послідовності значно відрізняються за довжиною. Якщо застосовувати основний алгоритм глобального вирівнювання, то отримуємо наступний результат:

AGACCAGATTTCTGCCAG
AGA_C _ _ _ _ _T_ _ _ _CAG

з вагою –12, на відміну від ваги попереднього вирівнювання, яка дорівнює –15. Друге вирівнювання, незважаючи на більшу вагу, буде гіршим, з точки зору пошуку співпадаючих фрагментів в послідовностях. Це відображує той факт, що, якщо не враховувати кінцеві пробіли, вага першого вирівнювання дорівнює 5. Алгоритм псевдоглобального вирівнювання застосовується з тими ж цілями, що і алгоритм локального вирівнювання, а саме для пошуку співпадаючих ділянок, а також для порівняння послідовностей, що дуже відрізняються по довжині.

Алгоритм псевдоглобального вирівнювання

1) *Побудова таблиці премій і штрафів p*.

2) *Ініціалізація масиву ваг вирівнювання префіксів a.*

Вирівнювання, коли не штрафується перший пробіл в *t*, еквівалентно найкращому вирівнюванню між *t* та підпослідовностями послідовності *s*. Значення $a(i, j)$ відповідає максимальній подібності префікса *t*, довжиною *i* та підпослідовностями префікса *s*, довжиною *j*. Ініціюючи нулями нульовий стовпчик, одержуємо випадок вирівнювання, коли ігноруються перші пробіли в послідовності *t*. Ініціюючи нулями нульові рядок та стовпчик одержуємо випадок вирівнювання, коли ігноруються перші пробіли в обох послідовностях.

$$\begin{matrix} a(0, j) = 0, \\ a(i,0) = 0, \end{matrix} \quad \text{де} \quad \begin{matrix} i = 0,...m, \\ j = 0,...n. \end{matrix}$$

3) *Розрахунок всіх елементів масиву a за рекурентною формулою (4.2).*

4) *Зворотній хід алгоритму*.

Розглянемо модифікацію основного алгоритму для побудови псевдоглобальнго вирівнювання, коли кінцевий пробіл в послідовності *t* не штрафується. Кінцевий пробіл в *t* співпадає з підпослідовностями послідовностей *s*. Ця частина вирівнювання – вирівнювання між *t* і префіксом *s*, і, щоб одержати оптимальне вирівнювання без врахування кінцевих пробілів, потрібно вирівняти *t* і префікс *s*. Подібність між *t* та префіксом *s* буде одержано, якщо взяти максимальне значення в останньому рядку масиву *a*, тобто $Max = \max\ a(m, j), \quad j = 1,...,n$ (таблиця 4.6).

Аналогічним чином потрібно діяти, якщо не штрафується кінцевий пробіл у *s*. Максимум в цьому випадку шукається в останньому стовпчику масиву *a* $Max = \max\ a(i, n), \quad i = 1,...,m$ (таблиця 4.6).

Якщо не штрафуються кінцеві пробіли у *t* та *s*, то для одержання вирівнювання потрібно знаходити максимальне значення в останніх рядку та стовпчику масиву *a*:

$$Max = \max\begin{cases} a(i,n), & i = 1,...m, \\ a(m,j), & j = 1,...n. \end{cases}$$

Значення *Max* дорівнює вазі найкращого псевдоглобального вирівнювання.

Таким чином, *зворотній хід алгоритму псевдоглобального вирівнювання* полягає у пошуку максимального елементу в останніх рядку та стовпчику масиву *a* і побудові зворотного шляху, рухаючись від зазначеного максимального значення масиву *a* (значення *Max*) до елементу з координатами $(0,0)$.

Таким чином, для вказаних послідовностей існує декілька значень *Max* = 1, які породжують три наступних варіанти вирівнювання:

_ _ CAAATC	CAAATC_ _	CAAATC
CT C_ _ _ _ _	_ _ _ _ _CTC	_ _ _CTC

i ↓	j→	*0*	*1*	*2*	*3*	*4*	*5*	*6*
	t \ *s*	_	**C**	**A**	**A**	**A**	**T**	**C**
0	_	0	←0	←0	←0	←0	←0	0
1	**C**	↑0	1	–1	–1	↖–1	–1	↖**1**
2	**T**	↑0	–1	0	–2	–2	↖0	–1
3	**C**	0	↖**1**	–1	–1	–3	–2	↖**1**

Таблиця 4.6. Псевдоглобальне вирівнювання: зворотній шлях вирівнювання для послідовностей *s* = CAAATC та *t* = CTC.

Очевидно, що ваги усіх трьох вирівнювань при ігноруванні пробілів рівні між собою $\omega = 1$. Це означає, що усі три вирівнювання є оптимальними. Ми розглянули приклад, у якому існує декілька максимальних елементів, але одне максимальне значення у останніх рядку і стовпчику також не рідкісний випадок.

4.4. Вирівнювання подібних послідовностей

Розглянемо алгоритм швидкого вирівнювання подібних послідовностей (лише для випадку глобального вирівнювання), коли порівнюються послідовності однакової довжини n. В цьому випадку масив $a(i,j)$ стає квадратною матрицею і по головній діагоналі розміщено вирівнювання без пробілів між s і t. Якщо таке вирівнювання не оптимальне, то необхідно ввести пробіли для одержання більшої ваги. Пробіли завжди вводяться попарно, один в s і один в t. Вирівнювання ведеться поблизу головної діагоналі.

Основна ідея полягає в тому, що у випадку схожих послідовностей, найкраще вирівнювання лежить в малому оточенні головної діагоналі. Розглянемо приклад алгоритму для заповнення матриці в горизонтальній та вертикальній смузі шириною $2k+1$ навколо головної діагоналі (рис.4.2).

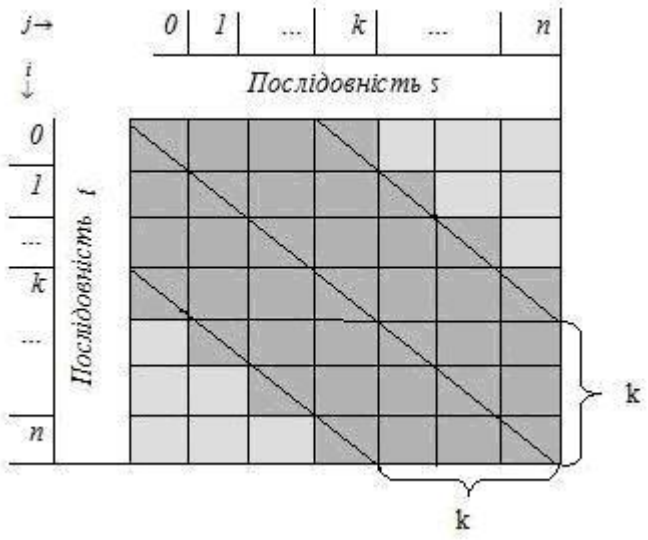


Рис. 4.2 Масив ваг вирівнювання префіксів a для порівняння подібних послідовностей однакової довжини

Елемент $a(n,n)$ містить максимальну вагу вирівнювання в смузі заданої ширини. Відмітимо, що значення поза смугою ширини k не використовуються і

відповідно не розраховуються. Для того, щоб визначити, чи лежить позиція (i, j) всередині смуги, використовується проста перевірка $-k \le i - j \le k$.

Як і в інших алгоритмах динамічного програмування, кожен елемент $a(i, j)$ залежить лише від трьох попередніх значень: $a(i-1, j), a(i-1, j-1), a(i, j-1)$. Але перевіряти елемент $a(i-1, j-1)$ не обов'язково, оскільки він знаходиться на тій самій діагоналі, що і елемент $a(i, j)$, і завжди буде знаходитися всередині смуги. Перевіряти потрібно елементи $a(i-1, j)$, $a(i, j-1)$, котрі можуть виходити за межі смуги, за умови того, що елемент $a(i, j)$ знаходиться всередині неї.

Модифікація основного алгоритму для вирівнювання подібних послідовностей

- Ініціалізація першого рядка і стовпчика масиву ваг префіксів *a*: перші *k* елементів рядка і стовпчика обчислюються за формулою:

$$\begin{matrix} a(0, j) = 0, \\ a(i,0) = 0, \end{matrix} \quad \text{де} \quad \begin{matrix} i = 0,...k, \\ j = 0,...k. \end{matrix}$$

(величина *k* наперед задана).

- Заповнення частини таблиці (рис 4.2), що знаходиться у смузі шириною $2k + 1$, за наступним принципом:

  Якщо елемент $a(i-1, j)$ знаходиться всередині смуги, тоді:

$$a(i, j) = \max \begin{cases} a(i-1, j) + g, \\ a(i-1, j-1) + p(i, j). \end{cases}$$

  Якщо елемент $a(i, j-1)$ знаходиться всередині смуги, тоді:

$$a(i,j)=\max\begin{cases} a(i,j-1)+g, \\ a(i-1,j-1)+p(i,j). \end{cases}$$

Очевидно, що у випадку коли обидва елементи $a(i-1,j)$ та $a(i,j-1)$ знаходяться всередині смуги, тоді елемент $a(i,j)$ потрібно розраховувати за вже відомою формулою (4.2).

Після першого проходження алгоритму, якщо одержане значення $a(n,n)$ для фіксованого значення k більше, або дорівнює вазі найкращого вирівнювання з k+1 пробілами, у такому випадку знайдено оптимальне вирівнювання. Враховуючи, що у вирівнюванні присутні принаймні k+1 пар пробілів, то можливо найкраща вага вирівнювання дорівнює $M\left[n-(k+1)\right]+2(k+1)g$. Ця величина вираховується з припущенням, що у вирівнюванні знаходиться точно k+1 пара пробілів, а усі інші пари дають співпадіння ($M>0$ – вага співпадіння).

Якщо $a(n,n)$ менше цієї величини, ширина смуги подвоюється і знову розраховується $a(n,n)$. Описаний метод можна розширити і для послідовностей різної довжини.

4.5. Загальна функція штрафу

Визначимо розрив в послідовності як безперервну кількість $k>1$ пробілів. З урахуванням виникнення можливих мутацій, поява розриву з k пробілами більш ймовірна, ніж поява k незалежних пробілів. Це пов’язано з тим, що поява довгого розриву може бути результатом лише однієї делеції, в той час як k пробілів, розділених блоками «символ-символ», утворюються через k мутацій, а виникнення однієї мутаційної події більш ймовірне, ніж декількох. Блочне вирівнювання враховує можливу кластеризацію пробілів.

Позначимо $w(k)$ функцію штрафу розриву з k пробілами. $w(k)=gk$, де g – величина штрафу за індивідуальний пробіл, k - кількість пробілів.

Відмінності між алгоритмом подібності послідовностей з урахуванням загальної функції штрафу та основним алгоритмом виникають через неадитивну схему штрафів. Кожне вирівнювання тепер можна розкласти на суму послідовних блоків трьох типів:

1) два символи;
2) максимальна серія послідовних символів в *s*, вирівняна з пробілами в *t*;
3) максимальна серія послідовних символів в *t*, вирівняна з пробілами в *s*.

Серія максимальна в тому сенсі, що не може бути продовжена далі. Розглянемо наступне блочне вирівнювання.

TTG_ _ _TTAAGGCTGATG
TGATCCA_ _ _ _ _ _ GCG_ _

T|T|G|_ _ _ |T|TAAGGC|TGA|TG
T|G|A|TGG|A|_ _ _ _ _ _|GCG|_ _

Блоки першого типу одержують вагу $p(i, j)$, де i та j – символи, що вирівнюються. Блоки другого та третього типів одержують вагу $w(k)$, де k – довжина розриву.

Тепер вага вирівнювання розраховується на рівні блоків. Адитивність ваг зберігається тільки на межах блоків, що призводить до деяких змін в основному алгоритмі. Блоки не можуть йти довільно один за одним, блоки другого та третього типів не можуть йти за блоками того ж типу. Для кожної пари (i, j) тепер зберігається не значення ваги найкращого вирівнювання між префіксом *t*, довжиною *i* та префіксом *s*, довжиною *j*, а вага вирівнювання між їх останніми блоками певного типу.

Для порівняння послідовності *t* довжиною *m* та послідовності *s* довжиною *n* використовують три масиви розміру $(m+1)(n+1)$:

Масив *a* – для порівняння символьних блоків;

Масив *b* – для порівняння блоків, що закінчуються пробілами в *t*;

Масив *c* – для порівняння блоків, що закінчуються пробілами в *s*.

Алгоритм подібності з урахуванням загальної функції штрафу

1) *Побудова таблиці премій і штрафів p*.

2) *Ініціалізація масивів ваг префіксів a,b,c.*

Масиви ініціюються за наступними формулами:

$$\begin{cases} a(i,0) = \omega(i), \\ b(i,0) = \omega(i), \\ c(i,0) = \omega(i). \end{cases} \qquad \begin{cases} a(0,j) = \omega(j), \\ b(0,j) = \omega(j), \\ c(0,j) = \omega(j). \end{cases}$$

3) *Розрахунок усіх інших елементів масивів a,b,c.*

В залежності від типу блоку, яким закінчується вирівнювання, змінюються значення у відповідних масивах за рекурентними формулами:

$$a(i,j) = p(i,j) + \max\begin{cases} a(i-1,j-1), \\ b(i-1,j-1), \\ c(i-1,j-1). \end{cases} \qquad b(i,j) = \max\begin{cases} a(i,j-k) + \omega(k), \\ c(i,j-k) + \omega(k), \\ 1 \le k \le j. \end{cases}$$

$$c(i,j) = \max\begin{cases} a(i-k,j) + \omega(k), \\ c(i-k,j) + \omega(k), \\ 1 \le k \le i. \end{cases}$$

Відмітимо, що значення масивів *b* і *c* залежать від декількох величин, оскільки останній блок може мати різну довжину.

4) *Пошук максимального з трьох елементів* $a(m,n)$, $b(m,n)$, $c(m,n)$ *та вибір стартової точки.*

Для одержання значення максимальної ваги вирівнювання береться максимум по всім $a(m,n)$, $b(m,n)$, $c(m,n)$.

5) Побудова вирівнювання на основі значень елементів масивів a, b, c.

Для вирівнювання використовується та ж ідея, що і в інших алгоритмах. Ми вибираємо елементи масивів з максимальною вагою, запам’ятовуючи масив якого типу було використано.

4.6 Оцінка статистичної значущості вирівнювання

Припустимо, що вирівнювання показує подібність двох послідовностей. Необхідно з’ясувати, чи має ця подібність біологічний зміст, чи співпадіння двох послідовностей є випадковим. Основні аспекти для з’ясування подібності послідовностей наступні:

- якого типу вирівнювання розглядаються;
- система оцінки якості вирівнювання (системи премій та штрафів, які н еобхідно використовувати);
- алгоритми, які використовуються для находження оптимальних вирівнювань;
- статистичні методі, які використовуються для оцінки значимости вирівнювання.

Таким чином статистичні методі оцінки значимости вирівнювання є не менш складною та важливою задачею для з’ясування подібності та біологічного змісту вирівнювання послідовностей ніж інші аспекти цього питання. Для деяких простих явищ таких, як підкидання монети або грального кубика реально, використовуючи статистичні методі, підрахувати розподіл результатів, що очікуються; для вирівнювання послідовностей ця задача нетривіальна.

Основний підхід до визначення значимості вирівнювань – це розрахунок **статистичної значимості ваги вирівнювань**, який базується на моделях Бернулі, моделях Маркова та їх модифікаціях. В біоінформатиці використовуються три основні величини для характеристики статистичної значущості вирівнювання – **Z-число** (Z-score), **E-число** (E-*value*) та **P-число** (P-*value*).

Основні етапи знаходження статистичної значимості ваги вирівнювань наступні:

1) Вирівнювання вибраної (фіксованої) послідовності з рандомізованими (випадковими) послідовностями багато разів (на ансамблі 100-500 рандомізованих послідовностей), відповідно знаходження оптимальних вирівнювань досліджуваної послідовності з кожною рандомізованою послідовностю w_i та зведення даних у таблицю.

При цьому рандомізовані (випадкові) послідовності можна отримати різними методами:

- Шляхом випадкової перестановки символів у досліджуваній послідовності без урахування порядку слідування символів у досліджуваній послідовності, наприклад, з використанням генератора випадкових чисел (модель Бернулі);
- Шляхом побудови рандомізованих послідовностей з урахуванням частот зустрічаємості пари та більшої кількості символів у послідовностях, тобто з урахуванням порядку слідування символів у послідовностях (моделі Маркова);
- Використання БД, як джерела великої кількості випадкових послідовностей, що в середньому мають незначну подібність з досліджуваною послідовнітю та ін.

Таким чином вирівнювання вибраної (фіксованої) послідовності з рандомізованими (випадковими) послідовностями отриманими шляхом випадкової перестановки символів у досліджуваній послідовності з використанням генератора випадкових чисел відповідають моделі Бернулі, якщо враховують тільки частоту зустрічаємості окремих символів у послідовностях. В моделях Маркова при побудові рандомізованих послідовностей враховуються статистичні властивості реальних амінокислотних послідовностей, наприклад, частоти зустрічаємості пари та більшої кількості символів, порядок слідування певної кількості символів,

сигнальні області (старт-кодони, стоп-кодони, сайти рестрикції, місця зв'язування РНК з ДНК, промотори і т.п.).

2) Побудова густини функції розподілу (або розподілу ймовірностей) ваг оптимальних вирівнювань між досліджуваною послідовностю та кожною рандомізованою послідовностю (рис.4.3).

Якщо $w \geq w_{\max}$, де $w_{\max}$ визначає положення максимуму густини функції розподілу ваг оптимальних вирівнювань, то зазначену густину функції розподілу можна наблизити формулою:

$$f(w_i) = 1 - \exp\left(-K\,\mathrm{e}^{-\lambda w_i}\right), \qquad (1)$$

де K і λ - параметри, пов'язані з розташуванням максимуму та шириною густини функції розподілу ваг оптимальних вирівнювань (рис. 1). При цьому права, повільно спадаюча частина графіку (далі «хвіст» густини функції розподілу ваг оптимальних вирівнювань, тобто область I на рис. 4.3), означає, що цей «хвіст» густини функції розподілу не описується нормальним розподілом:

$$f_0(w_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(w_i - \langle w \rangle\right)^2}{2\sigma^2}\right), \qquad (2)$$

в якому «хвіст» густини функції розподілу експоненційно швидко згасає до нуля. Тут σ^2 – дисперсія, w_i – вага вирівнювання досліджуваної послідовності з i-ою рандомізованою послідовністю, $\langle w \rangle$ – середнє значення ваг оптимальних вирівнювань досліджуваної послідовності з рандомізованими послідовностями. Таким чином, ваги в «хвості» густини функції розподілу (область I на рис. 4.3) є менш значущими (так як мають більшу ймовірність), ніж у випадку, коли б вони описувались нормальним розподілом (2). Проте формула (1) не описує частину графіку на рис. 4.3, що знаходиться ліворуч від положення максимуму

густини функції розподілу, так як не прямує до нуля при w_i що прямує до мінус нескінченності.

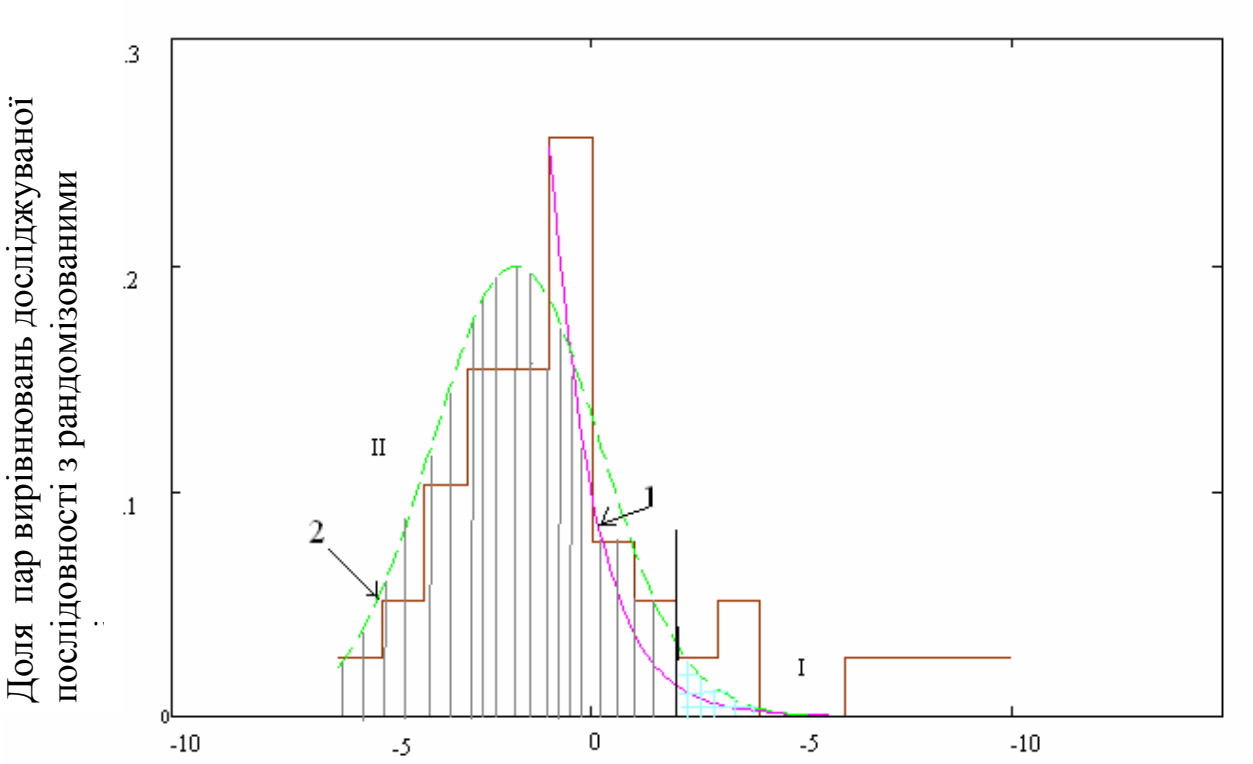


Рис. 4.3. Крива 1 – наближення розподілу ваг оптимальних вирівнювань між досліджуваною послідовністю та кожною рандомізованою послідовністю такої ж довжини за формулою нормального розподілу (2); крива 2 – наближення розподілу ваг оптимальних вирівнювань між досліджуваною послідовністю та кожною рандомізованою послідовністю такої ж довжини за формулою (1).

3) Оцінка значимості досліджуваного вирівнювання.

Як було відмічено, для оцінки значимості вирівнювання використовуються статистичні величини такі, як Z-число, P-число, E-число.

Z-число – це міра невипадковості співнадінь при вирівнюванні послідовностей. Розрахунок величини Z-числа відбувається за формулою:

$$Z = \frac{w - \langle w \rangle}{\sigma_w}, \qquad (3)$$

де w – вага вирівнювання досліджуваної послідовності з послідовністю з БД, або вага вирівнювання двох досліджуваних послідовностей;

$\langle w \rangle$ – середня вага оптимальних вирівнювань досліджуваної послідовності з рандомізованими послідовностями, яка дорівнює:

$$\langle w \rangle = \frac{\sum_{i=1}^{N} w_i n_i}{n}, \qquad (4)$$

N – кількість точок на графіку, σ_w – середньоквадратичне відхилення ваг вирівнювань досліджуваної послідовності з рандомізованими послідовностями, n_i - кількість рандомізованих послідовностей, що мають вагу w_i, n - загальна кількість рандомізованих послідовностей, при цьому

$$\sigma_w = \sqrt{\frac{\sum_{i=1}^{N} \left(w_i - \langle w \rangle \right)^2 n_i}{n(n-1)}}. \qquad (5)$$

З формули (3), що якщо Z = 0, то це означає, що розрахована подібність не краще ніж середня по ансамблю рандомізованих послідовностей і, ймовірно, що наявне співпадіння двох досліджуваних послідовностей є випадковим.

Статистичний аналіз вирівнювання двох послідовностей показав, що при $Z \geq 5$ співпадіння в оптимальному вирівнюванні двох послідовностей вважають значущим. Чим більше Z-число, тим більша ймовірність того, що вирівнювання не випадкове.

Крім Z-числа в якості міри значимості вирівнювань, як зазначено вище, використовується P-число та E-число. Більшість програм при роботі з БД розраховують P-число .

P-число – це ймовірність того, що знайдена подібність може бути випадковою, тобто ймовірність того, що досліджуване вирівнювання не краще ніж випадкове.

Орієнтовні значення P-числа та їх інтерпретації наступні:

$P \leq 10^{-100}$ – точне співпадіння;

$10^{-100} < P \leq 10^{-50}$ – послідовності майже ідентичні, наприклад, алелі або поліформізми;

$10^{-50} < P \leq 10^{-10}$ – гомологія очевидна, близькоспоріднені послідовності, близька гомологія;

$10^{-10} < P \leq 10^{-1}$ – скоріш за все дальноспоріднені послідовності, гомологія незначна, дальня гомологія;

$P > 10^{-1}$ – співпадіння не є значущим.

Деякі програми (серії BLAST) при роботі з БД (NCBI: http://ncbi.nlm.nih.gov, EMBL: http://ebi.ac.uk) розраховують E-число.

E-число – це очікувана кількість послідовностей в БД, що мають таке саме, або краще значення числа Z, що і досліджуване вирівнювання.

Орієнтовні значення E-числа та їх інтерпретації наступні:

$E \leq 0.02$ – послідовності ймовірно гомологічні;

$0{,}02 \leq E \leq 1$ – неможливо точно встановити гомологію, гомологія не очевидна;

$E > 1$ – випадкове співпадіння.

Ці два параметри (числа P та E) широко використовуються при порівнянні послідовностей, ця класифікація створена за досвітом фахівців. В кожному конкретному випадку обчислюється той параметр, яким зручніше користуватися. Наприклад, програма для порівняння послідовностей BLAST виводить E-числа. Це пов'язано з тим, що зручніше користуватися E-числами, які дорівнюють, наприклад, 5 і 10, ніж P-числами, які дорівнюють відповідно 0.993 і 0.99995. Однак, у випадку коли $E < 0.01$, то P-числа і E-числа є набагато меншими за одиницю, і E-число не є більш наочним для аналізу значимості, ніж P-число. Останні версії програм порівняння білка FASTA використовують саме цей підхід (EMBL: http://ebi.ac.uk).

4) Розрахунок відсотка ідентичних залишків.

Крім Р-числа і Е-числа програми для порівняння послідовностей видають відсоток ідентичних залишків. Існує безліч правил ітерпретації відсотка ідентичних залишків в оптимальному вирівнюванні. Розглянемо одно з них. Якщо два білки містять більше 45% ідентичних залишків в їх оптимальному вирівнюванні, то ці білки мають дуже схожі структури і, швидше за все, загальну або, принаймні, схожу функцію. Якщо вони містять від 25% до 45% ідентичних залишків, вони, ймовірно, мають схожий патерн фолдінга. З іншого боку, низька міра схожості послідовностей не може унеможливити гомології. Р. Ф. Дулітл визначив область 18%-25%-ої схожості послідовностей як «область двозначності», для якої можлива гомологія, але таке припущення може бути невірним. Парні вирівнювання, які знаходяться нижче за цю область (18%), мало, що можуть значити.

Хоча «область неоднозначності» ненадійна для висновків, але при прийнятті рішення про дійсну спорідненість важливо враховувати особливості вирівнювання: чи ізольовані схожі залишки і розподілені по всій послідовності або ж вони утворюють айсберги — локальні ділянки високої схожості (термін Р.Ф.Дулітла), які можуть відповідати загальному активному сайту, чи наявні загальні ліганди або функції. Якщо відомі просторові структури білків, то можливо перевірити їх схожість безпосередньо.

Ось декілька прикладів, які характерізують неоднозначність при прийнятті рішення про дійсну спорідненість вирівнювання.

-Міоглобін кашалота і леггемоглобін (червоний залізовмісний білок, по ряду властивостей схожий з гемоглобіном крові) люпину мають 15% ідентичних залишків в оптимальному вирівнюванні. Це навіть нижче області неоднозначності, визначену Р.Ф.Дулітлом. Але відомо, що обидві молекули мають схожі тривимірні структури, містять гемові простетичні (небілкові) групи і зв'язують кисень. Вони дійсно є віддаленими гомологами. Це приклад дивергенції (розходження признаків у видів, викликана різними умовами середовища).

-Якщо порівняти дві протеази: хімотрипсин і субтілізін, то Їх послідовності схожі на 12%. Це протеолітичні ферменти класу гідролаз, які розщеплюють пептидний зв'язок між амінокислотами в білках, виконують схожу функцію і в активному центрі несуть три характерні для них залишки. Проте ці ферменти мають різні просторові структури. Їх загальна функція і механізм – приклад конвергентної еволюції, тобто еволюційного процесу, який приводить до формування комплексу схожих признаків у представників неспоріднених груп.

Статистична значимість ваг вирівнювань — дуже важлива характеристика. Але при цьому треба чітко розуміти, що статистична значимість завжди оцінюється відносно випадкової моделі, а ці моделі можуть бути різними і мати різну міру адекватності реальності. Тому, приведені тут оцінки можуть носити орієнтовний характер, і не завжди означають реальну статистичну значимість.

ПРИКЛАД

Розглянемо знаходження статистичної значимості ваги вирівнювань на конкретному прикладі. Нехай S – досліджувана послідовність довжиною 48 символів.

1) Здійснимо вирівнювання досліджуваної (фіксованої) послідовності з рандомізованими (випадковими) послідовностями багато разів (на ансамблі 100 рандомізованих послідовностей) та знайдемо оптимальні вирівнювання досліджуваної послідовності з кожною рандомізованою послідовностю w_i, які отримані за допомогою генератора випадкових чисел на основі моделі Бернулі.

В якості досліджуваної послідовності візьмемо підпослідовність послідовності гемоглобіну людини:

S GGCTCGAGGA GCTCGTCTAG AGGATCGCTC GAGTGATCAG TCGGCCGC

M=48 (кількість символів у досліджуваній послідовності та послідовностях для порівняння)

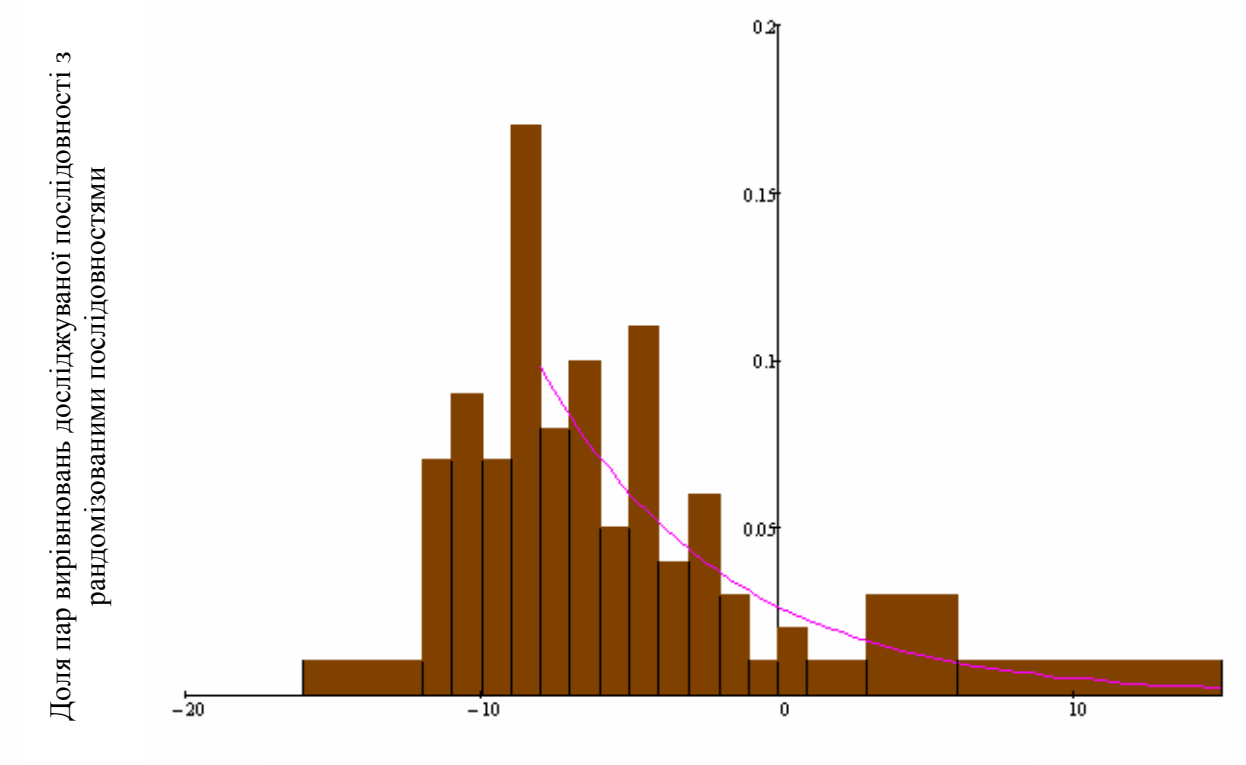
w_i = 38, 0, -3, -3, -4, -5, -11, -9, -9, -7, -7, -12, -10, -8, -11, -11, -8, -8, -9, -7, -16, -12, -12, -2, -10, -2, -10, -10, -7, -9, -9, -9, -11, -9, -8, -9, -7, -8, -9, -10, -10, -7, -9, -13, -10, -9, -11, -10, -9, -9, -9, -11, -12, -12, -11, -5, -11, -5, -8, -4, -5, -7, -5, -5, -9, -8, -7, -3, -8, -6, -5, -7, -12, -5, -7, -5, -6, -9, -4, -5, -6, -15, -6, -2, -4, -9, 3, 3, 3, -3, 0, 2, -3, -3, -12, -1, 1, 1, 6, -1, 11.

Дані вирівнювання представлені у таблиці 4.7 та у вигляді графіку на рисунку 2. Сумарна кількість рандомізованих послідовностей *n=100*.

Таблиця 4.7. Таблиця оптимальних вага вирівнювань пар послідовностей

n_i	-16	-15	-13	-12	-11	-10	-9	-8	-7	-6	-5
w_i	1	1	1	7	9	7	17	8	10	5	11
n_i	-4	-3	-2	-1	0	1	2	3	6	38	
w_i	4	6	3	1	2	1	1	3	1	1	

2) Побудуємо густину функції розподілу (або розподілу ймовірностей) оптимальних вирівнювань досліджуваної послідовності з кожною рандомізованою послідовностю (рис. 4.4).



Оптимальна вага вирівнювань пар

Рис. 4.4. 1 – гістограма експериментально отриманого розподілу ваг оптимальних вирівнювань між досліджуваною послідовністю та кожною рандомізованою послідовністю довжиною 48 символів на ансамблі із 100 рандомізованих послідовностей; 2 – наближення розподілу ваг оптимальних вирівнювань між досліджуваною послідовністю та кожною рандомізованою послідовністю такої ж довжини за формулою (1).

3) Оцінимо значимість досліджуваних вирівнювань.

Знаходимо вагу вирівнювання досліджуваної послідовності з послідовностями з БД (з досліджуваними послідовностями). Для прикладу обираємо нуклеотидну підпослідовність послідовності гемоглобіну жаби африканської S1 та свині S2.

S1: GGCCCTTGGC AGGCTGCTGA ATACTTTTCC CTGGACCCAA AGATACTTA

S2: GTAGTGAGCCG CGCGGCTCT AGAGAGCTA GCTAGTCTATC GGTCGTTCG

Оцінки значимості вирівнювань підпослідовності послідовності гемоглобіну людини і жаби наступні: середня вага $\langle w \rangle = -6{,}54$, середньоквадратичне відхилення $\sigma_w = 0{,}607$, вага оптимального вирівнювання двох досліджуваних підпослідовностей S і S1 w=-11, $Z = -7{,}35$, тобто співпадіння не є значущим.

Оцінки значимості вирівнювань підпослідовності послідовності гемоглобіну людини і свині наступні: середня вага $\langle w \rangle = -6{,}54$, середньоквадратичне відхилення $\sigma_w = 0{,}607$, вага оптимального вирівнювання двох досліджуваних підпослідовностей S і S2 w=6, $Z = 20{,}07$, тобто співпадіння не випадкове і є значущим.

4.7 Множинне вирівнювання послідовностей

У попередніх пунктах ми розглядали методи вирівнювання двох послідовностей, однак дуже часто необхідно дослідити подібність декількох послідовностей, наприклад, при необхідності дослідження послідовностей білків зі схожими функціями або для аналізу сімейства гомологічних білків. Для отримання такої інформації необхідно побудувати множинне вирівнювання послідовностей. За допомогою множинного вирівнювання можливо розпізнати слабку подібність, яка вважалася би незначущою при попарному порівнянні послідовностей. Однак така подібність може виявитися дуже значущою, якщо одні і ті самі залишки зберігаються і у інших, віддалено зв’язаних послідовностях. Множинне вирівнювання послідовностей відіграє надзвичайно важливу роль, як для вивчення білків, так і у генній інженерії та інших розділах біології. Зокрема, на його основі розраховуються матриці ваг амінокислотних замін, які потім використовуються при аналізі інших далеких гомологів.

Слід зазначити, що множинне вирівнювання – алгоритмічно складна задача. Для її вирішення найбільш поширені у використанні декілька алгоритмів:

1) Алгоритми, за допомогою яких можна знайти найкраще вирівнювання, для даної системи премій та штрафів. Такі алгоритми мають обмеження по кількості послідовностей і їх довжині;
2) Евристичні алгоритми, основані на попарному вирівнюванні;
3) Евристичні алгоритми, що будують глобальне вирівнювання, на основі локального вирівнювання;
4) Евристичні алгоритми, що будують локальне множинне вирівнювання.

Існують і інші методи множинного вирівнювання, такі як генетичні алгоритми, або методи, що основані на прихованих Марківських моделях, та інш.

Розглянемо множинне вирівнювання, що являє собою узагальнення вирівнювання двох послідовностей. Розглянемо k послідовностей, символи яких узяті з одного алфавіту $s_1, s_2, ... s_k$. Множинне вирівнювання таких послідовностей будується шляхом введення у них пробілів, таким чином, щоб їх довжини були рівні між собою. Зазвичай, послідовності розташовують одну над іншою, таким чином, щоб у стовпчиках можна було спостерігати відповідність знаків, або знаків і пробілів. Необхідною умовою побудови такого вирівнювання є те, що жоден стовпчик не повинен складатися виключно із пробілів. Найчастіше множинне вирівнювання використовується для білків, воно може виглядати наступним чином:

LVRIKKM
LLQ_KK_
LR_ IKKM
LPRVKKM.

Для визначення якості вирівнювання розглянемо спосіб знаходження ваги вирівнювання, що оснований на попарних вирівнюваннях. Обмежимося строго адитивними функціями, тобто будемо вважати, що вага вирівнювання дорівнює

сумі ваг стовпчиків. Для визначення ваги стовпчика необхідно визначити функцію з k аргументами (k – кількість послідовностей). Така функція повинна мати наступні властивості:

- не повинна залежати від порядку слідування аргументів. Наприклад, якщо один стовпчик містить знаки M,_,M,M,V, а інший – знаки M,M,V,_,M, то обидва стовпчики повинні отримати однакову вагу;
- повинна преміювати наявність багатьох подібних або генетично пов'язаних амінокислотних залишків і штрафувати незв'язані залишки та пробіли у кожному стовпчику.

Функція, що задовольняє цим умовам, називається *функція парних сум*. Вона визначається як сума парних ваг усіх пар символів у стовпчику

$$\omega_S(a,b,c,...) = p(a,b) + p(b,c) + p(a,c) + ...,$$

де $p(a,b)$ - парна вага символів a та b.

Наприклад, вага вищезазначеного стовпчика дорівнюватиме:

$$\omega_S(M,\_,M,M,V) = p(M,\_) + p(M,M) + p(M,M) + p(M,V) + \\ + p(\_,M) + p(\_,M) + p(\_,V) + p(M,M) + p(M,V) + p(M,V)$$

Завдяки простоті і ефективності, система парних сум використовується доволі часто. Слід зазначити, що у тому випадку, коли стовпчик містить декілька пробілів необхідно враховувати вагу $p(\_,\_) = 0$. Будемо позначати W_S як загальну вагу вирівнювання, її можна обчислити двома способами: або підрахувати ваги усіх стовпчиків, а потім додати їх, або підрахувати ваги усіх комбінацій попарних вирівнювань і також додати.

Як і у випадку вирівнювання двох послідовностей будемо використовувати метод динамічного програмування. Припустимо, що відомі k послідовностей, довжиною n: $s^1 = s_1^1 ... s_n^1$, $s^2 = s_1^2$, $...s_n^2...$, $s^k = s_1^k ... s_n^k$ (верхній індекс – номер послідовності, нижній індекс – номер символу у послідовності). У даному випадку для зберігання ваг префіксів знадобиться k-мірний масив a, розмірністю $(n+1)$ у кожному вимірі. Будемо використовувати лічильники $i_1, i_2, ..., i_n$, де $0 \le i_j \le n$, $j = 1,...n$.

Модифікація основного алгоритму для множинного вирівнювання

- *Ініціалізація масиву ваг префіксів a.*

$$a(0,...,0)=0.$$

- *Розрахунок всіх елементів масиву a.*

По аналогії з вирівнюванням двох послідовностей запишемо рекурентну формулу:

$$a(i,j)=\max\begin{cases} a(i_1-1,...i_n-1)+\omega_S\left(s^1_{i_1},...s^n_{i_n}\right), \\ a(i_1,i_2-1,...i_n-1)+\omega_S\left(\_,s^2_{i_2},...s^n_{i_n}\right), \\ a(i_1-1,i_2,i_3,...i_n-1)+\omega_S\left(s^1_{i_1},\_,s^3_{i_3},...s^n_{i_n}\right), \\ \ldots \\ a(i_1-1,...i_{n-1}-1,i_n)+\omega_S\left(s^1_{i_1},...s^{n-1}_{i_{n-1}},\_\right), \\ a(i_1,i_2,i_3-1,...i_n-1)+\omega_S\left(\_,\_,s^3_{i_3},...s^n_{i_n}\right), \\ \ldots \end{cases}$$

- *Побудова зворотнього шляху.*

Побудова зворотнього шляху починається з елементу $a(n,...n)$ і продовжується за аналогією з попарним вирівнюванням.

Проаналізуємо роботу алгоритму. Кожний елемент масиву залежить від 2^k-1 раніше обчислених елементів масиву. Оскільки необхідно вирівняти k послідовностей, то число можливих комбінацій у стовпчику дорівнює 2^k. Видаляючи заборонену комбінацію, що складається із пробілів, число комбінацій дорівнює 2^k-1, у це число входить множник 2^k, що має експоненційну часову складність. Експоненційна складність алгоритму метода динамічного програмування робить його використання неможливим для розв'язку широкого кола задач. Головну проблему створює розмір масиву.

Якщо використовувати стандартний підхід динамічного програмування, то розрахунок оптимальних множинних вирівнювань може зайняти довгий час,

незважаючи на скорочення числа оброблюваних комірок. Однак, як було зазначено раніше існують альтернативні евристичні методи, які простіші в реалізації, більш швидкі, але не дають гарантії якості знайденого вирівнювання, хоча у більшості випадках результат виявляється достатньо точним.

Один з таких методів – *вирівнювання зірки*. Основна ідея цього алгоритму полягає у тому, щоб по-перше, обрати послідовність, яка найбільш подібна до усіх інших послідовностей, а потім використати її як «центр зірки», вирівнюючи усі інші послідовності з нею. При розгляді даного методу будемо використовувати систему премій та штрафів (4.1).

Розглянемо наступні п’ять послідовностей ДНК.

s_1: ATTGCCATT

s_2: ATGGCCATT

s_3: ATCCAATTTT

s_4: ATCTTCTT

s_5: ACTGACC.

Спочатку необхідно розрахувати оптимальну вагу попарних вирівнювань (за допомогою алгоритму глобального вирівнювання двох послідовностей). Запишемо знайдені значення у матрицю і знайдемо суми кожного рядка (Таблиця 4.8).

	s_1	s_2	s_3	s_4	s_5	$\sum$*вага*
s_1	–	7	–2	0	–3	2
s_2	7	–	–2	0	–4	1
s_3	–2	–2	–	0	–7	–11
s_4	0	0	0	–	–3	–3
s_5	–3	–4	–7	–3	–	–17

Таблиця 4.8. Парні ваги послідовностей s_1, s_2, s_3, s_4, s_5.

З усіх послідовностей, послідовність s_1 має найбільшу сумарну вагу при вирівнюванні з іншими ($\omega = 2$), а отже саме цю послідовність потрібно обрати за центр майбутньої зірки. Запишемо попарні глобальні вирівнювання послідовності s_1 з усіма іншими послідовностями:

s_1: ATTGCCATT
s_2: ATGGCCATT,

s_1: ATTGCCATT
s_4: ATCTTC_ TT,

s_1: ATTGCCATT_ _
s_3: ATC_CAATTTT,

s_1: ATTGCCATT
s_5: ACTGACC_ _.

Починати побудову вирівнювання потрібно з одного з попарних вирівнювань і продовжувати, додаючи кожне наступне попарне вирівнювання. У ході цього процесу потрібно поступово додавати пробіли у центр зірки, щоб вони задовольняли наступним вирівнюванням і ніколи не видаляємо пробіли, які вже існують у центрі зірки, використовуючи так звану методику «один пробіл – завжди пробіл». У даному випадку, починаємо з s_2:

s_1: ATTGCCATT
s_2: ATGGCCATT.

Далі до цього вирівнювання потрібно додати s_3, оскільки s_3 довше ніж s_1 та s_2 потрібно додати кінцеві пробіли послідовностей s_1 та s_2:

s_1: ATTGCCATT_ _
s_2: ATGGCCATT_ _
s_3: ATC _CAATTTT.

Щоб отримати множинне вирівнювання потрібно ще додати s_4 та s_5, вони також повинні мати кінцеві пробіли.

s_1: ATTGCCATT_ _
s_2: ATGGCCATT_ _

s_3: ATC _CAATTTT

s_4: ATCTTC_ TT_ _

s_5: ACTGACC_ _ _ _.

Алгоритм вирівнювання зірки не завжди дає змогу знайти оптимальне множинне вирівнювання. Розглянемо приклад послідовностей, вирівнювання яких за допомогою описаного вище алгоритму не буде оптимальним.

s_1: GGCAA

s_2: GCACA

s_3: GGACA.

Центром зірки буде послідовність s_1, знайдемо оптимальні вирівнювання між нею та двома іншими послідовностями:

s_1: GGCA _ A $\qquad$ s_1: G_ GCAA

s_2: G _CACA, $\qquad$ s_3: GGACA _.

У результаті отримаємо наступне вирівнювання, вага якого, згідно міри попарних сум дорівнює $\omega_S = -5$.

s_1: GG _CA_A

s_2: G_ _CACA

s_3: GGACA _ _.

Однак для даних послідовностей оптимальним буде наступне множинне вирівнювання з вагою 0.

s_1: GG CA_A

s_2: G_ CACA

s_3: GG _ACA.

Запитання до розділу 4

1. Вкажіть основні причини необхідності порівняння біологічних послідовностей.
2. Що називають вирівнюванням послідовностей?
3. Яка величина характеризує якість вирівнювання?
4. Які основні особливості має алгоритм динамічного програмування?
5. Опишіть основні кроки алгоритму глобального вирівнювання.
6. У яких випадках застосування глобального вирівнювання дає суттєву інформацію для аналізу біологічних послідовностей?
7. Яке вирівнювання називається локальним? У яких випадках воно застосовується?
8. Охарактеризуйте відмінності між алгоритмами локального і глобального вирівнювання.
9. Опишіть основні кроки алгоритму псевдоглобального вирівнювання.
10. У чому полягає основна ідея алгоритму порівняння подібних послідовностей?
11. Які масиви використовуються для блочного порівняння послідовностей?
12. Опишіть характеристики величин, які використовуються для оцінки статистичної значущості вирівнювання?
13. Яке вирівнювання називається множинним? У яких випадках воно використовується?
14. Які методи множинного вирівнювання Вам віддомі?
15. Що таке функція парних сум?
16. У чому полягає складність алгоритму метода динамічного програмування для множинного вирівнювання?
17. Що таке вирівнювання зірки? Опишіть основні етапи цього алгоритму.

Тренувальні вправи до розділу 4

1. Нехай, відомі дві нуклеотидні послідовності s = ACTGTCCA, t = CTGAATCAGA. Використовуючи систему премій та штрафів (4.1), знайти вагу наступного вирівнювання:

   ACTG _ _T _ _CCA

   C _TGAAT _ CAGA.

   Чи буде таке вирівнювання оптимальним?

2. Знайти вагу наступних вирівнювань:

   AGTCACAAAC_ GCCAGAGGCTACACGATTCC_ _CTT _C_TGCAGG

   AGTCACAAAAAGACAAATACTGTATGATTTCAACTTACATC_AGG,

   GAGATTTATAGA_ AGCAGA

   AAG_ TCCATAGAGA_CAGA.

   Використовувати систему премій та штрафів (4.1).

3. Використовуючи систему премій та штрафів (4.1), знайти усі оптимальні глобальні вирівнювання нуклеотидних послідовностей s і t.

   а) s = CTTAGA, t = GTAA;

   б) s = AAGTTCGT, t = CAGTAAT.

4. Заповнити матрицю вирівнювання префіксів:

s / t	_	A	T	T	C	G	C
_	0	–1	–2	–3	–4	–5	–6
G	–1	0	–1	–2	–3	–3	–4
A	–2						
A	–3						
T	–4						
C	–5						
G	–6						
G	–7						
C	–8						

Використовувати наступну систему премій та штрафів:

- Однакові нуклеотиди +1;
- різні нуклеотиди 0;
- пробіли –1.

5. Розглянемо систему штрафів та премій для нуклеотидних послідовностей:
   - Однакові нуклеотиди +2;
   - різні нуклеотиди –3;
   - пробіли –4.

   Знайти усі оптимальні глобальні вирівнювання для наступних послідовностей: *s* = TGGCAAC, *t* = CTGGA.
6. Знайти оптимальне глобальне вирівнювання фрагментів ДНК миші s = AGGGTCCGGCGC, t = GAGGTCGAC. Систему штрафів та премій обрати самостійно.
7. Відомо, що кінь та кит плацентарні ссавці, а кенгуру сумчасте. Визначити, які з даних видів найбільш близькі, за допомогою послідовностей рібонуклеотиду коня (Equus caballus), малого полосатика (Bolaenoptera acutorostrata) та рудого кенгуру (Macropus rufus).

   а) Equus caballus

   KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTF
   VHEPLADVQAICLQKNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNC
   AYQTSQKERHIIVACEGNPYVPVHFDASVEVST

   в) Bolaenoptera acutorostrata

   RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVN
   TFVHESLEDVKAVCSQKNVLCKNGRTNCYESNSTMHITDCRQTGSSK
   YPNCAYKTSQKEKHIIVACEGNPYVPVHFDNSV

   с) Macropus rufus

   ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHE
   PKSVVDAVCHQENVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQY
   ETSNLNKQIIVACEGQYVPVHFDAYV

8. Нехай, відомі дві нуклеотидні послідовності s = AGTGGCATT, t = TGTCGCAT. Використовуючи наступні системи премій та штрафів знайти усі оптимальні локальні вирівнювання.

   а) Систему премій та штрафів (4.1);

   б) Однакові нуклеотиди +1;

   різні нуклеотиди 0;

   пробіли –1;

   в) Однакові нуклеотиди +5;

   різні нуклеотиди –4;

   пробіли –2.

9. Знайти оптимальне локальне вирівнювання фрагментів ДНК *Drosophila* s = ACATGCAAAAC, t = AAACT. Систему штрафів та премій обрати самостійно.
10. Відомі фрагменти ДНК послідовностей двох видів мавп. Фрагмент ДНК першої мавпи GGGCCATCCATGGGGGCATCC відповідає за синтез інсуліну, для другої мавпи відомий фрагмент кодуючої інсулін послідовності ATGGCCATGCACCC. Знайти інтронні зони фрагменту ДНК першої мавпи.
11. Знайти оптимальне псевдоглобальне вирівнювання фрагментів ДНК *Bifidobacterium infanis* s = GGATCACCTCCAACG, t = CCTAAGGCG. Систему штрафів та премій обрати самостійно.
12. Перетворіть алгоритм вирівнювання подібних послідовностей, таким чином, щоб він міг обробляти послідовності різної довжини.
13. Використовуючи вирівнювання зірки з системою премій та штрафів (4.1), побудувати множинне вирівнювання для наступних послідовнотей:

    s_1: AGTCCT

    s_2: ACTGTTC

    s_3: CCGCGTT

    s_4: GGGTCCT.

Література до розділу 4

1. Ж. Сетбуал, Ж. Мейданис Введение в вычислительную молекулярную биологию [пер. с англ. А.А. Чумичкина, под ред. А.А. Миронова], 2007, Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 420с.
2. A. Isaev Introduction to Mathematical Methods in Bioinformatics, Berlin, Springer, 2006, 294 p.
3. D.Higgins, W.Taylor Bioinformatics Sequence, structure and databanks, Oxford, University press, New York, 2001, 250 p.
4. A. M. Lesk Introduction to bioinformatics, Oxford University press New York, 2002, 255 p.
5. Д. Гасфильд Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология [пер. с англ. И.В. Романовского], СПб.: Невский Диалект, БХВ-Петербург, 2003. – 654 с.
6. М. С. Гельфанд, В. А. Любецки. Биоинформатика от эксперимента к компьютерному анализу и снова к эксперименту. Вестник РАН, том 73, № 11, с. 987-994 (2003)
7. Koonin E.V., Galperin M.Y. Sequence-Evolution-Function: Computational approachers in comparative genom-ics. Kluwer Academic Press, 2003.

РОЗДІЛ 5

МАТРИЦІ АМІНОКИСЛОТНИХ ЗАМІН

Емпіричні дослідження амінокислотних замін дозволили з'ясувати, що в процесі еволюції амінокислотні заміни відбувалися не рівномірно, амінокислоти частіше заміняються на подібні їм по фізико-хімічним властивостям, таким як розміру, гідрофобноті/гідрофільності, заряду, полярності та ін. (таблица 5.1). Так, наприклад, такі амінокислоти, як гліцин, цистеїн та триптофан заміняються рідко. Тому, якщо відбулася амінокислотна заміна, це може майже ніяк не вплинути на структуру та функцію білку, а може значно їх змінити. Так, якщо лізин заміниться на лейцин, який суттєво відрізняється від лізину, то просторова структура білка, а відповідно, і його функція може суттєво змінитися. А заміна лізину на аргінін може не вплинути на просторову структуру та функцію білка. Використовувати системи премій і штрафів для вирівнювання амінокислотних послідовностей, такі як +1 за збіг,-1 за неспівпадіння та -2 за пробіл, та подібні їм, неефективно. Тому при вирівнюванні амінокислотних послідовностей використовуються матриці амінокислотних замін для розрахунку ваг (або системи штрафів). Матриці амінокислотних замін використовуються також для дослідження еволюційних змін за будь який проміжок часу.

Характер амінокислотних замін визначається ступенем їх консервативності або радикальності. Консервативною заміною амінокислоти називається мутаційна заміна, яка не призводить до суттєвих змін структури та функції білка. В процесі еволюції консервативні заміни амінокислот відбуваються частіше, ніж радикальні. Ці заміни переважно зустрічаються в функціонально важливих ділянках білкової молекули (наприклад, сайтах зв'язування лігандів). Радикальні заміни амінокислот, напроти, суттєво змінюють структуру та функції білка.

Кольор	Тип залишку	Амінокислоти
Жовтий	Малий неполярний	Gly, Ala, Ser, Thr
Зелений	Гідрофобний	Cys, Val, Ile, Leu, Pro, Phe, Tyr, Met, Trp
Пурпурний	Полярний	Asn, Gln, His
Червоний	Негативно заряджений	Asp, Glu
Блакитний	Позитивно заряджений	Lys, Arg

Таблица 5.1. Класифікація амінокислот за їх фізико-хімічними властивостями

Існує багато параметрів, розрахунок яких дозволяє визначити характер замін амінокислот (фізико-хімічна дистанція Грентсема, індекс Танга, показник функціональної близькості амінокислот Бачінського, коефіцієнти Сніта та інші). Найбільш просто розраховується фізико-хімічна дистанція. Для її обчислення Р. Грентсем запропонував метод, який враховує вплив наступних трьох параметрів взаємозамінних амінокислотних залишків: полярність, об'єм та склад (відношення атомних мас невуглецевих атомів, які входять в кінцеві групи або кільця, до атомних мас вуглецевих атомів бокових радикалів).

Отримані Р. Грентсемом фізико-хімічні дистанції представлені в таблиці 5.2. При значенні фізико-хімічної дистанції більшому за 57,9, заміна вважається консервативною, в противному випадку – радикальною.

Відомо, що заміни амінокислот, які відбуваються в процесі еволюції можуть бути обумовлені заміною одного (однокрокова заміна), двох (двокрокова заміна) або трьох (трьохкрокова заміна) нуклеотидів в кодоні ДНК та відповідній іРНК. З'ясовано, що багатокрокові заміни амінокислот відбуваються в ході еволюції рідко, оскільки являються більш радикальними.

Всі заміни нуклеотидів можна розділити на 2 групи: синонімічні та несинонімічні. Синонімічні заміни – це заміни нуклеотидів, які не призводять до зміни амінокислоти що кодується. Несинонімічні заміні – це заміни нуклеотидів, в результаті яких відбувається зміна амінокислот, що кодуються.

	A	C	D	E	F	G	H	I	K	L	M	N	P	O	R	S	T	V	W	Y
A	100	9	41	50	45	73	59	54	50	54	59	45	86	54	45	54	73	68	32	45
C		100	27	23	4	27	18	9	4	9	9	36	23	27	18	45	32	9	0	9
D			100	77	18	54	59	23	50	18	27	86	50	68	54	68	59	27	14	27
E				100	36	54	82	36	73	36	41	77	54	86	73	64	68	41	27	41
F					100	27	54	86	50	86	86	27	45	45	54	27	50	77	82	86
G						100	54	36	41	36	41	64	77	59	41	73	73	50	14	32
H							100	54	82	54	59	68	64	86	86	59	77	59	45	59
I								100	50	95	95	32	54	50	54	32	59	86	68	82
K									100	50	54	54	50	73	86	41	64	54	50	59
L										100	91	27	54	45	50	32	54	82	68	82
M											100	32	59	50	54	36	59	86	68	82
N												100	54	77	59	77	68	36	18	32
P													100	64	50	64	82	68	32	50
O														100	77	68	77	54	41	54
R															100	50	64	54	50	64
S																100	73	41	18	32
T																	100	68	41	54
V																		100	59	73
W																			100	82
Y																				100

Таблиця 5.2. Матриця фізико-хімічних дистанцій Р. Грентсема для амінокислотних замін.

5.1 Розрахунок матриці амінокислотних замін.

Метод Дейхофф.

Для врахування цих факторів Маргарет Дейхофф в 1978 р. запропонувала використовувати так звані PAM-матриці (Point Accepted Mutation або Percent Accepted Mutation: точкові реалізовані мутації або відсоток реалізованих мутацій) для розрахунку ймовірності заміни однієї амінокислоти іншою. Для побудови матриці PAM потрібно розрахувати матрицю ймовірності мутацій *M* та важливе сімейство вагових матриць *S*. Ці матриці використовуються для розрахунку:

- ваг (або системи штрафів) при вирівнюванні амінокислотних послідовностей (матриці S)
- для дослідження еволюційних змін за будь який проміжок часу (матриці M).

Кожен елемент M_{ab} цієї матриці (таблиця 5.3) заміщень означає ймовірність, з якою амінокислота в рядку a заміняється на амінокислоту в стовпчику b за одиницю часу еволюції. За одиницю часу в матриці прийнято середній час, за який відбувається одна заміна амінокислоти в 100 сайтах. Дэйхофф запропонувала вимірювати кількість замін амінокислот у PAM (point accepted mutations, крапкові зафіксовані мутації, 1 PAM - 1 заміна амінокислоти на 100 сайтів). Тому матриця замін амінокислот Дэйхофф часто називається PAM-матрицею, на підставі якої Дэйхофф запропонувала свою одиницю швидкості еволюції білка - число PAMів, що накопичилося за 100 млн. років. За допомогою матриці (M) Дэйхофф можна передбачати еволюційні зміни амінокислот за будь-який період часу, якщо відомо початкову амінокислотну послідовність.

Нехай є дві вирівняні послідовності s і t (довжиною N). Наша задача з'ясувати наскільки значущими є співпадіння, які спостерігаються. Ми будемо розглядати мутації як не спрямовані події, тобто, для пари амінокислот a, b, що вирівняли, неважливо відбулася заміна $a \rightarrow b$ чи $b \rightarrow a$. Припустимо, що на i-ій позиції послідовності s знаходиться символ x_i, а на i-ій позиції послідовності t – символ y_i. Ймовірність знайти символ x_i на i-ій позиції послідовності s дорівнює

$$P(x_i) = N_{x_i} / N \qquad (5.1)$$

де N_{x_i} – кількість появ символу x_i в послідовності s.

Амінокислота		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ала	Арг	Асн	Асп	Цис	Глн	Глу	Гли	Гис	Иле	Лей	Лиз	Мет	Фен	Про	Сер	Тре	Три	Тир	Вал
A	Ала	9867	1	4	6	1	3	10	21	1	2	3	2	1	1	13	28	22	0	1	13
R	Арг	2	9913	1	0	1	9	0	1	8	2	1	37	1	1	5	11	2	2	0	2
N	Асн	9	1	9822	42	0	4	7	12	18	3	3	25	0	1	2	34	13	0	3	1
D	Асп	10	0	36	9859	0	5	56	11	3	1	0	6	0	0	1	7	4	0	0	1
C	Цис	3	1	0	0	9973	0	0	1	1	2	0	0	0	0	1	11	1	0	3	3
Q	Глн	8	10	4	6	0	9876	35	3	20	1	6	12	2	0	8	4	3	0	0	2
E	Глу	17	0	6	53	0	27	9865	7	1	2	1	7	0	0	3	6	2	0	1	2
G	Гли	21	0	6	6	0	1	4	9935	0	0	1	2	0	1	2	16	2	0	0	3
H	Гис	2	10	21	4	1	23	2	1	9912	0	4	2	0	2	5	2	1	0	4	3
I	Иле	6	3	3	1	1	1	3	0	0	9872	22	4	5	8	1	2	11	0	1	57
L	Лей	4	1	1	0	0	3	1	1	1	9	9947	1	8	6	2	1	2	0	1	11
K	Лиз	2	19	13	3	0	6	4	2	1	2	2	9926	4	0	2	7	8	0	0	1
M	Мет	6	4	0	0	0	4	1	1	0	12	45	20	9874	4	1	4	6	0	0	17
F	Фен	2	1	1	0	0	0	0	1	2	7	13	0	1	9946	1	3	1	1	21	1
P	Про	22	4	2	1	1	6	3	3	3	0	3	3	0	0	9926	17	5	0	0	3
S	Сер	35	6	20	5	5	2	4	21	1	1	1	8	1	2	12	9840	32	1	1	2
T	Тре	32	1	9	3	1	2	2	3	1	7	3	11	2	1	4	38	9871	0	1	10
W	Три	0	8	1	0	0	0	0	0	1	0	4	0	0	3	0	5	0	9976	2	0
Y	Тир	2	0	4	0	3	0	1	0	4	1	2	1	0	28	0	2	2	1	9945	2
V	Вал	18	1	1	1	2	1	2	5	1	33	15	1	4	0	2	2	9	0	1	9901

Таблиця 5.3. Матриця амінокислотних замін Дейхофф для еволюційної дистанції в 1 PAM.

**Примітка. Усі значення для зручності помножені на 10000.*

Аналогічно:

$$P(y_i) = N_{y_i} / N, \qquad (5.2)$$

де N_{y_i} – кількість появ символу y_i в послідовності t. У випадку, коли вирівнювання є випадковим, ймовірність зустріти на i-ій позиції послідовності s символ x_i, а на i-ій позиції послідовності t символ y_i дорівнює добутку $P(x_i)P(y_i)$. Тоді повна ймовірність вирівнювання

$$P_{\text{вип}}(s,t) = \prod_{i=1}^{N} P(x_i) \prod_{i=1}^{N} P(y_i) = \prod_{i=1}^{N} P(x_i)P(y_i). \tag{5.3}$$

Коли ж вирівнювання не є випадковим, ймовірність знайти символ x_i проти символу y_i не дорівнює добутку ймовірностей знаходження цих символів в кожній з послідовностей, тобто

$$P(x_i, y_i) \neq P(x_i)P(y_i). \tag{5.4}$$

В цьому випадку обчислимо ймовірність $P(x_i, y_i)$ як відношення кількості пар символів (x_i, y_i) до загальної кількості пар символів у вирівняних послідовностях. Тоді повна ймовірність вирівнювання

$$P(s,t) = \prod_{i=1}^{N} P(x_i, y_i). \tag{5.5}$$

Відношення ймовірності даного вирівнювання до ймовірності випадкового вирівнювання дорівнює

$$\frac{P(s,t)}{P_{\text{вип}}(s,t)} = \frac{\prod_{i=1}^{N} P(x_i, y_i)}{\prod_{i=1}^{N} P(x_i)P(y_i)} \tag{5.6}$$

і характеризує закономірність наявності співпадінь між послідовностями.

Прологарифмуємо це відношення, в результаті отримаємо:

$$\lg\left\{\frac{\prod_{i=1}^{N} P(x_i, y_i)}{\prod_{i=1}^{N} P(x_i)P(y_i)}\right\} = \sum_{i=1}^{N} \lg\left\{\frac{P(x_i, y_i)}{P(x_i)P(y_i)}\right\} = \sum_{i=1}^{N} S(x_i, y_i), \qquad (5.7)$$

де

$$S(x_i, y_i) = \lg\left\{\frac{P(x_i, y_i)}{P(x_i)P(y_i)}\right\}. \qquad (5.8)$$

Матриця S отримала назву вагова матриця елементів x_i та y_i. Якщо вирівнювання є випадковим (тобто $P(x_i, y_i) = P(x_i)P(y_i)$), то

$$S(x_i, y_i) = 0. \qquad (5.9)$$

Чим більше $S(x_i, y_i)$ відрізняється від 0, тим більш значущим є конкретне вирівнювання.

З формул (5.7) і (5.8) одержимо

$$\frac{P(s,t)}{P_{\text{вип}}(s,t)} = \prod_{i=1}^{N} \frac{P(x_i, y_i)}{P(x_i)P(y_i)} = \prod_{i=1}^{N} 10^{S(x_i, y_i)} = 10^{\sum_{i=1}^{N} S(x_i, y_i)}. \qquad (5.10)$$

Таким чином матриця S визначає наскільки значущими є співпадіння між двома послідовностями.

Визначимо матрицю 1 PAM як матрицю M_{ab}, що характеризує такий час (еволюційний період), на протязі якого очікується одна заміна в поліпептидному ланцюзі середнього складу зі 100 амінокислот. Далі введемо поняття частоти дозволених мутацій f_{ab}, яка дорівнює числу замін елементу a на елемент b. З огляду на те, що в даній моделі мутація розглядається, як неспрямована подія, можна записати

$$f_{ab} = f_{ba}. \tag{5.11}$$

Сумарна кількість мутацій по амінокислоті a дорівнює

$$f_a = \sum_{b=1}^{20} f_{ab}, \tag{5.12}$$

а сумарна кількість амінокислот, що мутували (подвоєна кількість мутацій) визначається співвідношенням

$$f = \sum_{a=1}^{20} f_a. \tag{5.13}$$

Тепер за частотами f_{ab} і ймовірностям *P(a)* можна побудувати 1 PAM. $P(a)$ – це ймовірності зустріти амінокислоту a на наборі білків (сімейств білків) на основі яких розраховується матриця PAM. Для прикладу в таблиці 5.4 наведено ймовірності (частоти) появи амінокислотних залишків $P(a)$ для білків, що занесені в БД NCBI.

Ala (A) 7,64%	Leu (L) 9,55%
Arg (R) 5,2%	Lys (K) 5,96%
Asn (N) 4,34%	Met (M) 2,36%
Asp (D) 5,25%	Phe (F) 4,1%
Cys (C) 1,62%	Pro (P) 4,89%
Gln (Q) 3,94%	Ser (S) 7,07%
Glu (E) 6,4 7%	Thr (T) 5,56%
Gly (G) 6,87%	Trp (W) 1,21%
His (H) 2,25%	Tyr (Y) 3,15%
Ile (I) 5,85%	Val (V) 6,62%

Таблиця 5.4 Ймовірності (частоти) появи амінокислотних залишків $P(a)$ в БД NCBI.

Матриця 1PAM розміром 20x20 з елементами M_{ab}, що дорівнюють ймовірності амінокислотної заміни *a* на *b*:

$$M_{ab} = \frac{f_{ab}}{100 fP(a)} \tag{5.14}$$

Відмітимо, що у випадку *a*=*b*, M_{aa} відповідає ймовірності відсутності заміни у певний проміжок часу для амінокислоти a. Для PAM матриць M_{aa} розраховується за відносною мутабільністю. Мутабільність – це ймовірність того, що за певний проміжок часу амінокислота *a* буде замінена на будь-яку амінокислоту.

$$m_a = \frac{f_a}{100 fP(a)}. \tag{5.15}$$

Тому ймовірність для амінокислоти *a* залишитись без змін дорівнює

$$M_{aa} = 1 - m_a. \tag{5.16}$$

Відмітимо, що всі розрахунки проводяться для спрощеної моделі еволюції білка. Припускається, що амінокислота *a* мутує незалежно як від еволюційної історії білку, так і від сусідніх амінокислот, які в принципі могли б впливати на мутабільність *a*.

Легко показати, що матриця ймовірності мутацій має наступні властивості:

$$1) \sum_{b=1}^{20} M_{ab} = 1,$$

$$2) \sum_{a=1}^{20} p_a M_{aa} = 0{,}99. \tag{5.17}$$

Перше рівняння прямо вказує на те, що для *a* сума ймовірності заміни на іншу амінокислоту і ймовірності залишитись без змін дорівнює 1. Нагадаємо,

що ці ймовірності відносяться до певного еволюційного періоду. Одиницю еволюції можна уявити собі як еволюційний період, за який в середньому на 100 амінокислот відбувається одна заміна. Це і є еволюційна відстань в 1 PAM. Матриця ймовірностей переходів нормована таким чином, щоб враховувати цей факт. Якби при розрахунку відносної мутабільності m_a в знаменнику було би використано, наприклад 50 замість 100, то властивості матриці не змінилися б, але одиниця еволюції тепер вказувала на період, за який в середньому вже на 50 амінокислот відбувається 1 заміна.

За матрицею M можна розрахувати ймовірності переходів для різних еволюційних відстаней. Наприклад, яка ймовірність того, що b замінить a через два еволюційних періоди? За перший період a може перейти в c з ймовірностю M_{ac}, за другий с може перейти в b з ймовірностю M_{cb}. Складаючи, одержимо, що ця ймовірність дорівнює M^2_{ab}. В загальному випадку M^k це матриця перехідних ймовірностей для k еволюційних періодів. Слід відмітити, що при зростанні k (порядку тисячі) матриця M збігається до матриці з однаковими стовпчиками, кожен з яких містить відносні ймовірності появи b, p_b. Тобто, незалежно від "початкового наближення", після досить довгого еволюційного проміжку з ймовірностю p_b, будь-яка амінокислота перейде в b.

Вагова матриця визначається наступним чином. Її значеннями є відношенням ймовірності мутації $a \rightarrow b$ до ймовірності випадкової появи b, інакше кажучи, відношення правдоподібності M_{ab}/p_b. На практиці використовується десятинний логарифм відношення. Ми розглядали вагову матрицю для відстані в 1 PAM, що легко може бути узагальнено на випадок k. Вагова матриця для відстані в k PAM визначається як

$$S_k(a,b) = \lg\left\{\frac{M^k_{ab}}{P(b)}\right\}. \qquad (5.18)$$

Матриця симетрична внаслідок того, що $f_{ab}=f_{ba}$.

Нехай $k = 1$, тоді

$$S(a,b) = \lg\left\{\frac{M_{ab}}{P(b)}\right\} = \lg\left\{\frac{P(a,b)}{P(a)P(b)}\right\}, \qquad (5.19)$$

тобто вагова матриця PAM при такому значенні k співпадає з матрицею замін.

Незважаючи на те, що матриці PAM розраховуються тільки на наборах дозволених мутацій і частот зустрічаємості амінокислот, вони відображають загальні властивості амінокислот, наприклад такі, як гідрофобність/гідрофільність, розмір, заряд, полярність та ін. Звичайно амінокислоти з подібними властивостями мають високу ймовірність взаємозамін, що не дивно хоча б тому, що схожі амінокислоти частіше беруть участь в дозволених мутаціях. Заміни за принципом подібності дозволяють зберегтися функції білку.

Найчастіше для двох послідовностей нічого не відомо про еволюційну відстань. В таких випадках рекомендується порівнювати послідовності з використанням різних матриць відстаней, наприклад 40 PAM, 120 PAM і 250 PAM. Ці матриці відрізняються тим, що побудовані для різних еволюційних відстаней. В загальному випадку менші PAM використовуються для пошуку коротких ділянок повної гомології, а більші – для виявлення довгих частково гомологічних фрагментів.

5.2 Матриці блочних замін BLOSUM

Матриці амінокислотних замін PAM з’явилися у той час, коли доступні дані головним чином складалися з сімейств послідовностей близько зв’язаних білків. Для таких послідовностей можна було оцінити лише короткострокову еволюційну матрицю, а потім за допомогою екстраполяції отримати матрицю для більш довгих еволюційних відстаней. Пізніше, коли стали доступні дані для більш віддалено зв’язаних білків, виявилося, що матриці PAM не в змозі відобразити дійсні довгострокові амінокислотні заміни.

Подружжя Хенікофф розробили сімейство матриць BLOSUM (BLOcks SUbstitution Matrix) для штрафів замін при порівнянні амінокислотних послідовностей. Їх метою було замінити матриці Маргарет Дейхофф. Варто зазначити, що матриці PAM розраховуються за принципом глобального вирівнювання, а матриці BLOSUM – локального (тільки висококонсервативних ділянок). Перевагою матриць BLOSUM є те, що отримують її з прямих даних, а не з екстрапольованих значень, як у випадку матриць PAM.

Матриці BLOSUM позначаються як BLOSUMr, де $1 < r < 100$. На відміну від матриць PAM, значення r обернено пропорційне еволюційній відстані, наприклад BLOSUM80 застосовується для коротких еволюційних відстаней, а BLOSUM45 – для довгострокових еволюцій. BLOSUM62 – найпоширеніша матриця сімейства BLOSUM. Значення елементів цієї матриці відображають заміни амінокислот, які були помічені у великій вибірці (більше 2000) консервативних регулярних комбінацій амінокислот, названих блоками. Ці блоки були знайдені в базі даних білкових послідовностей, що містить більше 500 сімейств родинних білків. BLOSUM62 часто використовується у тому випадку, коли необхідно отримати вирівнювання, що не має містити проміжків. Матриці BLOSUM50 розраховані для вирівнювань, коли проміжки допустимі.

При середній подібності послідовностей найбільш часто використовуються матриці PAM 160 і BLOSUM 62 які можна співставити між собою (Таблиця 5.5).

PAM	100	120	160	200	250
BLOSUM	90	80	60	52	45

Таблиця 5.5. Співставлення матриць PAM та BLOSUM

На рис. 5.1 та рис. 5.2 зображені приклади матриць PAM 250 та BLOSUM 45, які використовуються для аналізу послідовностей, що мають однакову міру подібності.

PAM 250

	C	G	P	S	A	T	D	E	N	Q	H	K	R	V	M	I	L	F	Y	W
C	12																			
G	-3	5																		
P	-3	-1	6																	
S	0	1	1	1																
A	-2	1	1	1	2															
T	-2	0	0	1	1	3														
D	-5	1	-1	0	0	0	4													
E	-5	0	-1	0	0	0	3	4												
N	-4	0	-1	1	0	0	2	1	2											
Q	-5	-1	0	-1	0	-1	2	2	1	4										
H	-3	-2	0	-1	-1	-1	1	1	2	3	6									
K	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
R	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
V	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
M	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
I	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
L	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
F	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Y	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
W	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17

Рис.5.1 Матриця PAM 250.

BLOSUM 45

	G	P	D	E	N	H	Q	K	R	S	T	A	M	V	I	L	F	Y	W	C
G	7																			
P	-2	9																		
D	-1	-1	7																	
E	-2	0	2	6																
N	0	-2	2	0	6															
H	-2	-2	0	0	1	10														
Q	-2	-1	0	2	0	1	6													
K	-2	-1	0	1	0	-1	1	5												
R	-2	-2	-1	0	0	0	1	3	7											
S	0	-1	0	0	1	-1	0	-1	-1	4										
T	-2	-1	-1	-1	0	-2	-1	-1	-1	2	5									
A	0	-1	-2	-1	-1	-2	-1	-1	-2	1	0	5								
M	-2	-2	-3	-2	-2	0	0	-1	-1	-2	-1	-1	6							
V	-3	-3	-3	-3	-3	-3	-3	-2	-2	-1	0	0	1	5						
I	-4	-2	-4	-3	-2	-3	-2	-3	-3	-2	-1	-1	2	3	5					
L	-3	-3	-3	-2	-3	-2	-2	-3	-2	-3	-1	-1	2	1	2	5				
F	-3	-3	-4	-3	-2	-2	-4	-3	-2	-2	-1	-2	0	0	0	1	8			
Y	-3	-3	-2	-2	-2	2	-1	-1	-1	-2	-1	-2	0	-1	0	0	3	8		
W	-2	-3	-4	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-3	-2	-2	1	3	15	
C	-3	-4	-3	-3	-2	-3	-3	-3	-3	-1	-1	-1	-2	-1	-3	-2	-2	-3	-5	12

Рис.5.2 Матриця BLOSUM 45.

Матриці замін широко використовуються у так званому алгоритмі BLAST (рис. 5.3). Програма «BLAST» (BLAST - Basic Local Alignment Search Tool - основний (програмний) засіб пошуку локальних вирівнювань) була написана у 1990 році. Ця програма спочатку шукає ідентичність більш коротких ділянок послідовностей, а потім намагається розширити їх у будь-якому напрямку, для того, щоб знайти більш довге вирівнювання. Така стратегія виправдана з біологічної точки зору. Користуватися BLAST можна за допомогою доступного Internet ресурсу http://blast.ncbi.nlm.nih.gov/Blast.cgi.

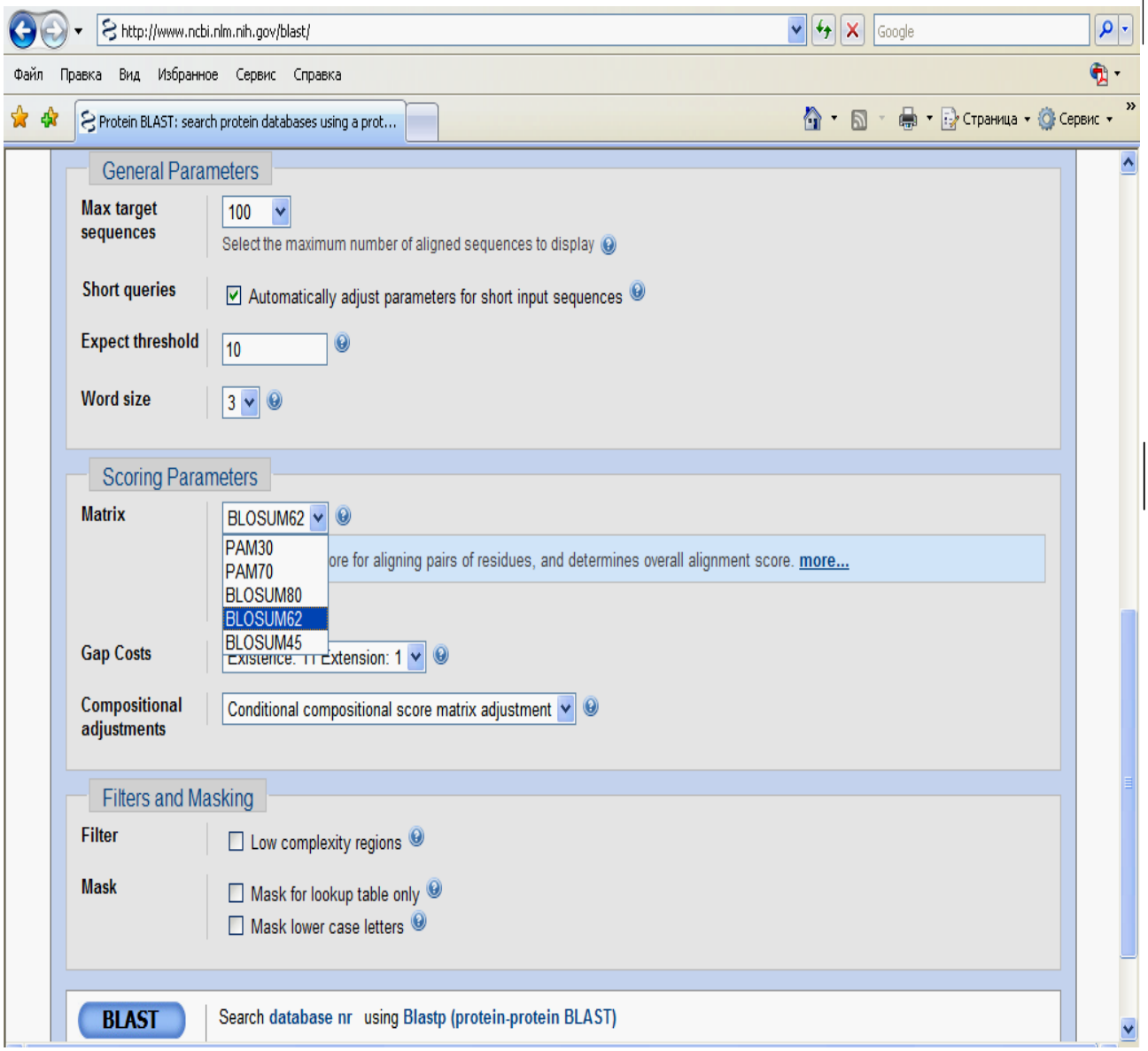


Рис.5.3 Використання матриць замін у алгоритмі BLAST.

5.3. Методи вивчення еволюційних змін амінокислотних послідовностей

Основним параметром, що характеризує еволюцію амінокислотних послідовностей, є еволюційна дистанція між ними. Еволюційні дистанції використовуються для побудови філогенетичних дерев (дендрограм) та визначення часу дивергенції, що можливо завдяки існуванню однозначної відповідності між часом дивергенції та ступенем амінокислотних відмінностей при умові постійності швидкостей еволюції білків в різних філогенетичних лініях.

Всі методи обчислення еволюційних дистанцій між амінокислотними послідовностями можна розділити на три групи:

1. Методи приблизної оцінки еволюційних дистанцій (кількість амінокислотних відмінностей, p-дистанція),
2. Методи коректованої оцінки еволюційних дистанцій (Пуассон-коректована дистанція, дистанція з поправкою Кімури, гама-дистанція, дистанція Грішина і т.д.),
3. Методи, які базуються на використанні матриць замін амінокислот (Дейхофф та Джонса-Тейлора-Торнтона).

Розглянемо кожну з цих груп більш детально.

Методи приблизної оцінки еволюційних дистанцій

Кількість амінокислотних відмінностей. Вивчення еволюційних змін білків та поліпептидів починається з порівняння та вирівнювання двох або більше амінокислотних послідовностей різних організмів.

Як вже зазначалося, вирівнюванням амінокислотних послідовностей називається процес співставлення порівнюваних послідовностей для такого їх взаєморозташування, при якому спостерігається максимальна кількість збігів амінокислотних залишків. До програм, призначених для вирівнювання амінокислотних послідовностей, відносяться Clustal W Protein та Multalin Protein, а також програмні пакети, які їх містять (наприклад, MEGA).

Найбільш простим методом визначення еволюційних відмінностей після проведення вирівнювання є підрахунок кількості амінокислотних відмінностей n_{d} між двома послідовностями. Однак для використання цього параметра необхідно, щоб послідовності, які порівнюються, складались з однакової кількості амінокислотних залишків n, що спостерігається досить рідко. На практиці амінокислотні послідовності часто включають пробіли, які не повинні враховуватися при обчисленні n_{d}.

Таким чином аналіз еволюції амінокислотних послідовностей по кількості відмінностей є можливим лише при однаковій довжині послідовностей. В інших випадках цей параметр є лише проміжною величиною для подальших розрахунків.

P-дистанція. Більш точним методом визначення кількості відмінностей між послідовностями є обчислення долі різних амінокислот в цих послідовностях (позначається p, P_d, *p-distance* від англ. part of differences). Використовуючи цю характеристику, можна оцінити кількість відмінностей між послідовностями навіть, якщо їх довжини значно відрізняються. p-дистанція визначається за формулою:

$$p_d = n_{\mathrm{d}} / n. \tag{5.20}$$

Якщо заміни у всіх амінокислотних сайтах відбуваються з рівною ймовірністю однієї заміни на сайт p_1, то ймовірність k замін на сайт розподілена за біноміальним розподілом:

$$P(k) = C_n^k p_1^k (1 - p_1)^{n-k}, \tag{5.21}$$

де $C_n^k = \dfrac{n!}{k!(n-k)!}$, k - кількість замін.

Дисперсія величини $P(k)$ визначається за формулою:

$$D(p) = np(1 - p). \tag{5.22}$$

Слід відмітити, що *p*-дистанція не є строго пропорційною часу дивергенції таксономічних груп організмів *t* і тому описаний метод використовується для отримання приблизної оцінки еволюційних відмінностей амінокислотних послідовностей (таксономічна група – група організмів, що об'єднані певним ступенем спорідненості, спільними рисами будови та функціональних особливостей).

Методи коректованої оцінки еволюційних дистанцій

Пуассон-коректована дистанція. Однією з причин нелінійної залежності *p*-дистанції від *t* є поступове зростання відмінності між n_d та істинною кількістю замін амінокислот при множинних замінах на певних ділянках сайтів. Для більш точного визначення кількості замін слід використовувати корекцію Пуассона.

Позначимо швидкість амінокислотних замін в рік в певному сайті *r* і будемо для спрощення вважати, що *r* у всіх сайтах однакова. Далі ми покажемо, що похибка, яка з'являється в наслідок такого припущення, є малою при малих значеннях $p_d (p_d \ll 1)$. Добуток *rt* дорівнює середній кількості амінокислотних замін на сайт за період часу *t*. Нехай ймовірність *P*(*k*) здійснення *k* амінокислотних замін в сайті описується розподілом Пуассона з інтенсивністю *rt*. Тоді

$$P(k) = \frac{(rt)^k}{k!} e^{-rt}. \qquad (5.23)$$

Отже, ймовірність відсутності амінокислотних замін в сайті дорівнює

$$P(0) = e^{-rt}, \qquad (5.24)$$

а очікувана кількість консервативних (незмінних) амінокислотних сайтів дорівнює ne^{-rt}. На практиці амінокислотна предкова послідовність, як правило, невідома, що робить неможливим безпосереднє використання формули (5.4), яка описує розподіл Пуассона. Кількість амінокислотних замін оцінюється при порівнянні двох гомологічних послідовностей, які дивергували t років назад. Оскільки ймовірність відсутності заміни в амінокислотному сайті послідовності дорівнює e^{-rt}, то ймовірність q того, що ні в одному з гомологічних сайтів двох послідовностей не відбудеться заміни, дорівнює

$$q = (e^{-rt})^2 = e^{-2rt}. \tag{5.25}$$

Ймовірність того, що заміни відбудуться можна записати як

$$p = 1 - q. \tag{5.26}$$

Отримана формула (5.5) для q є наближеною, оскільки не враховує зворотні та паралельні мутації (однакові мутації, які відбуваються в гомологічних амінокислотних сайтах в двох різних еволюційних лініях). Однак внесок цих ефектів є малим в широкому діапазоні значень $p \geq 0{,}3$.

Використовуючи отримані результати знаходимо, що загальна кількість амінокислотних замін на сайт $d = 2rt$ для двох послідовностей дорівнює:

$$d = -\ln(1 - p). \tag{5.27}$$

Величина d (5.7) має назву Пуассон-коректованої дистанції (Poisson correction distance, PC-distance).

Якщо, використовуючи додаткові джерела вдається визначити час дивергенції двох послідовностей одна від одної, то швидкість замін амінокислот дорівнює

$$r = d/2t. \tag{5.28}$$

Знаходження в знаменнику величини $2t$ (а не t) викликано тим, що t – це час, що пройшов після еволюційної дивергенції двох ланцюгів від спільної для

них предкового ланцюга і множник 2 в знаменнику відповідає двом гілкам філогенетичного дерева.

Дистанція Кімури. Еволюційна дистанція Кімури dK є ще одним варіантом корекції p-дистанції. Її величина обчислюється по емпіричній формулі

$$d_K = -\ln(1 - p - 1/5p^2). \quad (5.29)$$

Кімура з'ясував, що ця формула справедлива при значеннях p, які не перевищують 0,7.

Позначимо вектор-рядок відносних частот появи амінокислот поліпептиду в момент часу t_0 як p_0. Тоді частоти замін амінокислот за час t (або для t PAMів) визначаються формулою

$$p_0 = p_t M^t, \quad (5.30)$$

Елемент M^t_{ab} матриці дорівнює ймовірності того, що амінокислота в рядку i заміниться на амінокислоту в стовпчику b за час t *PAM*, а M^t_{aa} – ймовірності незмінності амінокислоти a за проміжок часу t *PAM*. Ці ймовірності можна використати для співставлення долі різних амінокислот у гомологічних послідовностях p та числом амінокислотних замін на сайт (d_D – дистанція Дейхофф), припустивши, що частоти амінокислот залишаються незмінними в процесі еволюції. Тоді p обчислюється як

$$p = 1 - \sum_a p(a) M_{aa}^{2t}, \quad (5.31)$$

де $p(a)$– це частота появи амінокислоти в послідовності. Використання M_{aa}^{2t} замість M_{aa}^{t} пов'язано з тим, що ми розглядаємо дві послідовності, які дивергували t *PAM* часових одиниць тому назад.

Не зважаючи на те, що для різних білків частоти появи амінокислот g відрізняються, Дейхофф запропонувала використовувати середні частоти

появи амінокислот для багатьох різних білків. Такий підхід не враховує специфічність кожного білка, але при цьому робить можливим застосування цього методу для багатьох білків. До того ж велика кількість білків має досить близькі частоти появи амінокислот, що робить цей метод досить точним.

Метод Джонса-Тейлора-Торнтона. У 1992 р. Джонс, Тейлор і Торнтон запропонували для побудови матриці замін новий метод (Jones-Taylor-Thornton matrix, JTT-matrix), який враховує більшу кількість замін для більшого числа білків. Також були зроблені спроби створення окремих матриць для мітохондріальних білків хребетних.

Оскільки різні білки мають різні матриці замін, було б доречно створити матриці замін для кожної групи білків, але для цього необхідно більше даних по їх амінокислотним послідовностям. Тому зараз для вимірювання еволюційної дистанції між послідовностями визначають параметр *a* та обчислюють відповідну гама-дистанцію.

Запитання до розділу 5

1. Які методи обчислення еволюційних дистанцій Вам відомі?
2. Охарактеризуйте особливості методу обчислення еволюційних дистанцій Дейхофф.
3. На чому базуються процедура розрахунку ймовірності заміни однієї амінокислоти іншою?
4. Опишіть основні етапи розрахунку матриці PAM.
5. Які властивості має матриця ймовірності мутацій?
6. У чому полягає перевага матриць BLOSUM?

Тренувальні вправи до розділу 5

1. Побудувати матрицю амінокислотних замін, використовуючи більш подібні послідовності задачі №7 (розділ 5). Порівняти її з матрицею Дейхофф. Побудувати вагову матрицю для 250 еволюційних дистанцій, порівняти отриману матрицю з матрицею PAM250.
2. Побудувати вагову матрицю для 1 PAM еволюційних дистанцій:
   - використовуючи більш подібні послідовності задачі №1 (розділ 5);
   - використовуючи той самий набір амінокислот, але переставлених у послідовності випадковим чином.

   Порівняти отримані результати.
3. Вирівняти амінокислотні послідовності за допомогою системи премій та штрафів (5.1) і за допомогою матриці PAM 250. Порівняти отримані результати.
4. Розрахувати час роботи алгоритму вирівнювання для двох послідовностей довжинами по 1000 та по 10 000 символів. При розв'язанні задачі необхідно приймати до уваги, що час роботи алгоритму $\tau = Cnm$, де $C = const$; n, m - кількість символів у послідовностях. При розв'язанні задачі виконувати наступні кроки:

- Виміряти τ_{20} - час вирівнювання послідовностей із 20 символів;
- Розрахувати C;
- Виміряти τ_{100} - час вирівнювання послідовностей із 100 символів;
- Розрахувати τ_{100}, і порівняти отримані результати.
- Розрахувати τ_{1000}, τ_{10000}.

5. Виміряти час вирівнювання амінокислотних послідовностей задачі №7 (розділ 5) за допомогою власної програми та за допомогою програми BLAST. Порівняти отримані результати та зробити висновки.

Література до розділу 5

1. Дурбин Р., Эдди Ш., Крог А., Митчисон Г. Анализ биологических последовательностей. – М. – Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2006. – 480 с.
2. Сетубал Ж., Мейданис Ж. Введение в вычислительную и молекулярную биологию. – М. – Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. – 420 с.
3. Игнасимуту С. Основы биоинформатики. – М. – Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. – 320 с.
4. Гладков Л.А., Курейчик В.В., Курейчик В.М. Генетические алгоритмы / Под. ред. В.М. Курейчика. – 2-е изд. испр. и доп. – М.: ФИЗМАТЛИТ, 2006. – 320 с.
5. Каменская М.А. Информационная биология – М.: Издательский центр “Академия”, 2006.
6. Глазко В.И., Глазко Г.В. Введение в генетику, биоинформатика, ДНК-технология, генная терапия, ДНК-экология, протеомика, метаболика – К.:КВІЦ, 2003.
7. Lesk M. Introduction to Bioinformatics – Oxford University Press Inc. New York, 2002.
8. Албертс Б. Молекулярная биология клетки: в 3 т. – М.: Мир, 1994.
9. Bioinformatics: Sequence, structure and database – Oxford University Press, 2001.

РОЗДІЛ 6

ТЕОРІЯ ЙМОВІРНОСТІ ЯК ІНСТРУМЕНТ БІОІНФОРМАТИКИ

6. 1. Поняття випадкової величини та ймовірності

Всі події, які відбуваються навколо нас, можна розділити на дві групи. До однієї групи відносяться події, які при реалізації певних умов, обов'язково відбудуться. До іншої групи належать події, результати яких не визначаються однозначно умовами спостережень. Наприклад, підкидаючи монету, ми не знаємо, якою стороною вона впаде; стріляючи однотипними снарядами без зміни наводки, в одну точку влучити неможливо. Виконуючи повторні високоточні вимірювання, зазвичай, отримують лише приблизно рівні, але різні результати. Такі події називаються випадковими.

Випадковою величиною називається величина, що в результаті експерименту може прийняти певне значення, яке не визначається однозначно умовами експерименту. Випадкові величини можуть бути дискретними та неперервними. *Дискретною випадковою величиною* називається величина, яка приймає дискретний набір значень (тобто може приймати окремі значення). Неперервні випадкові величини – це величини, які приймають значення з певного інтервалу. У біоінформатиці працюють з дискретними випадковими величинами.

Незалежні випадкові події – це події, для яких здійснення однієї з них ніяк не впливає на здійснення інших. У якості прикладу можна навести випадіння орла або решки при багаторазовому підкиданні монети. Аналогічним чином визначаються незалежні випадкові величини.

Для кількісного опису можливості здійснення випадкової події вводять поняття ймовірності. Найбільш просто визначається ймовірність у так званій елементарній теорії ймовірностей. Нехай при проведенні експерименту вимірюється деяка випадкова величина, значення якої залежить від реалізації

тієї або іншої випадкові події. Якщо зазначені випадкові події A_1, A_2, A_3, … задовольняють наступним умовам:

– в результаті проведення експерименту реалізується одна і тільки одна з цих подій;

– ці події є рівноможливими,

тоді ймовірність того, що випадкова величина прийме певне значення буде дорівнювати відношенню кількості сприйнятливих можливостей отримання результатів досліду до загальноможливої кількості цих результатів.

Це можна записати наступним чином:

$$P = m/n, \tag{6.1}$$

де m – кількість подій A_1, A_2, …, які сприятливі для набуття випадковою величиною обраного значення; n – загальна кількість можливих випадків. Співвідношення (6.1) називають класичним визначенням ймовірності.

Визначати ймовірність подій теоретично складно, а часто і неможливо. В зв'язку з цим широко використовується статистичне визначення ймовірності. Згідно цього визначення ймовірність – це величина, яка дорівнює відношенню кількості випадків, в яких певна подія спостерігалась до загальної кількості спостережень. Таку ймовірність ще називають відносною частотою. При цьому заздалегідь невідомо, скільки буде успішних дослідів і скільки їх буде всього. Для прикладу, розглянемо підкидання монети. Випадковою подією в даному випадку є випадіння орла (або решки). Якщо ми повторимо дослід (підкидання) багато разів, то випадкові події випадіння орла в кожному випробуванні будуть незалежними. Теоретична ймовірність випадіння орла дорівнює ½. Якщо підкидати монету і обраховувати відносну частоту, то вона не співпаде з теоретичною ймовірністю. Але при досить великій кількості випробувань величина відносної частоти випадіння орла буде як завгодно близько наближатись до теоретичної ймовірності цієї події.

Використовуючи статистичне визначення ймовірності можна записати:

$$P = N_{\text{р}}/N \qquad (6.2),$$

де $N_{\text{р}}$ – кількість випадків випадіння орла, а N – загальна кількість випробувань. Подія, ймовірність якої дорівнює 1, називається достовірною ($P = 1$), а подія, ймовірність якої дорівнює 0, називається неможливою ($P = 0$).

Розглянемо наступний приклад. Відомо, що в процесі біосинтезу білку в поліпептидний ланцюг включається 20 амінокислот.

Нехай P_A – ймовірність зустріти в послідовності амінокислоту A. Для того, щоб знайти P_A, необхідно записати амінокислотну послідовність білка (взявши її, наприклад, з бази даних) та визначити скільки разів в ній зустрічається дана амінокислота. Згідно з формулою (6.2), ймовірність P_A дорівнює

$$P_{\text{A}} = N_{\text{A}}/N, \qquad (6.3),$$

де N_{A} – кількість випадків, в яких в послідовності зустрічається амінокислота А, N – кількість всіх амінокислот в даній послідовності. При цьому

$$\sum_{\text{i}=1}^{20} P_{\text{i}} = 1. \qquad (6.4)$$

6.2. Числові характеристиками випадкової величини

У багатьох випадках, щоб описати випадкову величину з точки зору теорії ймовірностей, можна обмежитись лише окремими числовими параметрами, що характеризують найбільш суттєві властивості. Ці величини називаються числовими характеристиками випадкової величини.

Математичне сподівання випадкової величини характеризує середнє значення випадкової величини і визначається (для дискретних величин) наступним чином:

$$M(x) = \sum_i x_i p_i \,, \qquad (6.5),$$

де x_i – можливі значення випадкової величини, а p_i – їх ймовірність.

Зміст математичного сподівання $M(x)$ – середнє значення випадкової величини.

Математичне сподівання має наступні елементарні властивості.

1. Якщо $c = const$, то $M(c) = c$.
2. Нехай X – деяка випадкова величина, тоді $M(X + c) = M(X) + M(c)$.
3. $M(c \bullet X) = c \bullet M(X)$.

Величина $[X – M(X)]$ має назву центрованої випадкової величини.

Дисперсія випадкової величини характеризує ступінь розсіяння значень випадкової величини відносно її математичного очікування і дорівнює математичному очікуванню квадрата центрованої випадкової величини:

$$D(x) = M\left\{\left[x - M(x)\right]^2\right\}. \qquad (6.6)$$

Дисперсія має наступні найпростіші властивості.

1. Якщо $c = const$, то $D(c) = 0$.
2. Нехай X – деяка випадкова величина, тоді $D(X + c) = D(X)$.
3. $D(c \bullet X) = c^2 \bullet D(X)$

Використовуючи властивості математичного сподівання, вираз для дисперсії можна перетворити до більш зручного вигляду:

$$D(x) = M\left\{\left[x - M(x)\right]^2\right\} = M\left\{x^2 - 2xM(x) + M^2(x)\right\} =$$

$$= M(x^2) - 2M(x)M(x) + M^2(x) = M(x^2) - M^2(x)\,. \qquad (6.7)$$

Таким чином

$$D(x) = M(X^2) - (M(X))^2 . \qquad (6.8)$$

Дисперсія випадкової величини має розмірність квадрату випадкової величини, що є не дуже зручним. Такого недоліку позбавлено середньоквадратичне відхилення, розмірність якого співпадає з розмірністю самої величини. Середньоквадратичне відхилення випадкової величини X характеризує ширину діапазону значень X та визначається наступним чином

$$\sigma(X) = \sqrt{D(X)} . \qquad (6.9)$$

Математичне сподівання і дисперсія – найбільш розповсюдженні характеристики випадкової величини. Вони характеризують найважливіші властивості розподілу: найбільш ймовірне значення випадкової величини та ступінь розсіяння її значень.

6.3. Закони розподілу випадкової величини

Для опису випадкових величин зручно користуватись законом розподілу. *Закон розподілу випадкової величини* – це співвідношення між значеннями випадкової величини та ймовірностями їх реалізації. Він може задаватися таблицею, формулою або графіком.

Наприклад, закон розподілу для випадіння орла або решки при підкиданні монети можна зобразити у вигляді:

Результат	Орел	Решка
Ймовірність	½	½

Для прикладу з визначенням ймовірності знаходження амінокислоти в поліпептидному ланцюгу закон розподілу можна зобразити за допомогою таблиці:

G	A	V	……..	P
P_G	P_A	P_V		P_P

Для неперервних випадкових величин неможливо задати закон розподілу у вигляді таблиці. Для опису розподілу ймовірностей таких величин вводять поняття функцій розподілу. Функція розподілу випадкової величини $F(x)$ описує залежність ймовірності того, що випадкова величина X у випробуванні прийме значення менше x. Схематичний вигляд функції розподілу наведений на рис. 6.1.

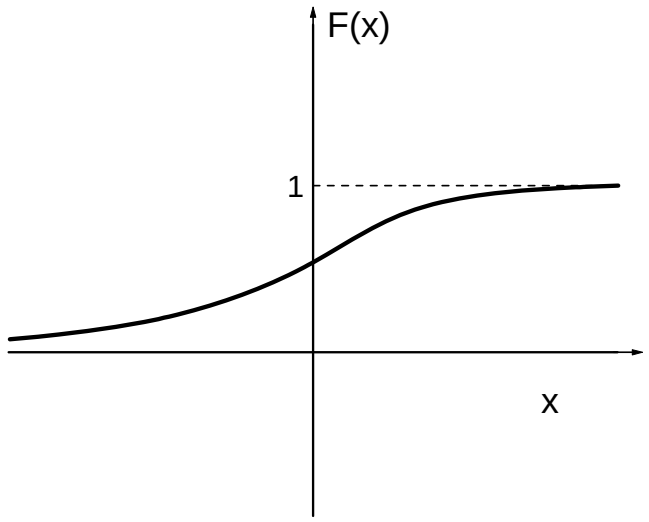


Рис.6.1.

Легко бачити, що функція розподілу *F(x)* – монотонно зростаюча функція. При цьому

якщо $x \to -\infty$, то $F(x) = 0$;

якщо $x \to \infty$, то $F(\text{x}) = 1$.

Знаючи функцію розподілу, можна легко визначити ймовірність того, що випадкова величина потрапить у заданий інтервал $a < X < b$. Вона визначається співвідношенням:

$$P(a < X < b) = F(b) - F(a).$$

В багатьох випадках для опису випадкової величини зручно використовувати не функцію розподілу, а пов'язану з нею густину розподілу ймовірностей, яка визначається наступним чином:

$$f(x) = \frac{dF}{dx}. \qquad (6.10),$$

При цьому

$$P(a \le X \le b) = \int_a^b f(x)dx. \qquad (6.11)$$

Справді, використовуючи визначення густини розподілу ймовірностей, отримаємо:

$$\int_a^b f(x)dx = \int_a^b \frac{dF(x)}{dx}dx = F(x)\Big|_a^b = F(b) - F(a) = P(a < X < b). \qquad (6.12)$$

З формули (6.12) очевидним чином випливає співвідношення:

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

6.3.1 Біноміальний розподіл

Широкого розповсюдження набув біноміальний розподіл. Розподіл такого типу виникає в схемі незалежних випробувань Бернуллі. Розглянемо цю схему більш детально.

Схемою Бернуллі називають послідовність незалежних в сукупності експериментів, в кожному з яких можливі лише два результати – "успіх" або "невдача"; при цьому "успіх" в одному експерименті наступає з ймовірністю p, а "невдача" – з ймовірністю $q = 1 - p$. Позначимо через $P_k(n)$ ймовірність того, що в n експериментах буде досягнуто k успіхів. Тоді, для будь якого $k = 0,1,2,\ldots, n$, має місце формула:

$$P_k(n) = C_n^k \cdot p^k (1-p)^{n-k}, \qquad (6.13)$$

де $C_n^k = \dfrac{n!}{k!(n-k)!}$ – біноміальні коефіцієнти. Доведемо це співвідношення.

Нехай проводиться n незалежних випробувань, в результаті кожного з яких може наступити або не наступити деяка подія А ("успіх"). За умовами проведення експериментів ймовірність здійснення події А дорівнює $P(\text{A}) = p$ і, таким чином, ймовірність протилежної події ("невдача") дорівнює $P(\overline{\text{A}}) = 1 - p = q$. Зазначимо, що здійснення або нездійснення події А може чергуватися різним чином. Будемо записувати можливі результати випробувань у вигляді комбінацій букв А і $\overline{\text{A}}$. Наприклад, запис $\text{A}\,\overline{\text{A}}\,\overline{\text{A}}\,\text{A}$ означає, що в чотирьох випробуваннях подія здійснилась у 1-му і 4-му випадках і не здійснилась у 2-му і 3-му.

Будь-яку комбінацію, у яку А входить k разів і $\overline{\text{A}}$ входить n - k разів, назвемо сприятливою. Кількість сприятливих комбінацій дорівнює числу способів, якими можна вибрати k однакових елементів з n елементів (такий

спосіб вибору елементів називають сполученням з n елементів по k), тобто $m = C_n^k$.

Розрахуємо ймовірність появи сприятливої комбінації. Розглянемо спочатку випадок, коли подія А відбувається у перших k випробуваннях і не відбувається в решті $n - k$ випробуваннях. Така комбінація має наступний вигляд:

$$\mathrm{B}_1 = \underbrace{\mathrm{AA...A}}_{k \text{ разів}}\,\underbrace{\overline{\mathrm{A}}\,\overline{\mathrm{A}}...\overline{\mathrm{A}}}_{n-k \text{ разів}}$$

Ймовірність появи цієї комбінації внаслідок незалежності випробувань (теорема множення ймовірностей) дорівнює:

$$P(\mathrm{B}_1) = \underbrace{P(\mathrm{A})P(\mathrm{A})...P(\mathrm{A})}_{k \text{ разів}}\,\underbrace{P(\overline{\mathrm{A}})P(\overline{\mathrm{A}})...P(\overline{\mathrm{A}})}_{n-k \text{ разів}} = p^k q^{n-k}$$

В будь-якій іншій сприятливій комбінації B_i подія А зустрічається також k разів, а подія $\overline{\mathrm{A}}$ відбувається $n - k$ разів, то ймовірність кожної з таких комбінацій також дорівнює $p^k q^{n-k}$. Отже

$$P(\mathrm{B}_1)=P(\mathrm{B}_2)=...=P(\mathrm{B}_\mathrm{m}) = p^k q^{n-k}. \qquad (6.14)$$

Всі розглянуті комбінації є, очевидно, несумісними. Тому (за аксіомою складання ймовірностей)

$$P_k(n) = P(\mathrm{B}_1+\mathrm{B}_2+...+\mathrm{B}_\mathrm{m}) = P(\mathrm{B}_1)+P(\mathrm{B}_2)+...+P(\mathrm{B}_\mathrm{m}) = mp^k q^{n-k} = C_n^k \cdot p^k \left(1-p\right)^{n-k}$$

У багатьох випадках виникає необхідність обчислити ймовірність певного числа успіхів в схемі Бернуллі при великій кількості експериментів (тобто, коли фактично $n \to \infty$) та малій ймовірності успіху. Однак при цьому безпосередньо використовувати формулу біноміального розподілу стає досить складно.

Зазначені складнощі пов'язані з тим, що при p, яке не прямує до нуля, ймовірність отримати будь-яке кінцеве число успіхів при необмеженому збільшенні числа експериментів ($n \to \infty$) прямує до нуля. Тому, в цьому випадку, ймовірність успіху буде залежати від кількості експериментів в серії: якщо експеримент один – ймовірність успіху p_1, якщо експериментів в серії два – ймовірність успіху p_2, …, якщо експериментів в серії n – ймовірність успіху p_n. Ймовірність успіху змінюється не всередині однієї серії (у відповідності з означенням схеми Бернуллі), а від серії до серії, коли змінюється загальна кількість експериментів.

6.3.2. Розподіл Пуассона

Введемо параметр $\lambda = np \to \lambda > 0$ та виконаємо в формулі для біноміального розподілу граничний перехід, покладаючи $n \to \infty$. Отримаємо:

$$C_n^k \cdot p^k (1-p)^{n-k} = C_n^k \cdot \frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k} =$$

$$= \frac{n(n-1)...(n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1-\frac{\lambda}{n}\right)^n \cdot \left(1-\frac{\lambda}{n}\right)^{-k}. \qquad (6.15)$$

При $n \to \infty$ кожен з чотирьох множників в останньому виразі веде себе по різному. Перший множник буде прямувати до 1. Справді, при $n \to \infty$ чисельник прямує до n^k, внаслідок того, що в кожному з k множників порівняно з n можна знехтувати всім іншим. Третій множник є визначною границею і прямує до $e^{-\lambda}$. Четвертий множник буде прямувати до одиниці. Враховуючи вище сказане, можна записати, що

$$\text{при } n \to \infty \;\; C_n^k \cdot p^k (1-p)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

Таким чином, при великих n має місце формула:

$$P_k(n) = \frac{\lambda^k}{k!} e^{-\lambda}. \qquad (6.16)$$

Цей розподіл отримав назву розподілу Пуассона, а параметр λ – інтенсивності розподілу.

Залишилось нез'ясованим важливе питання: в якому випадку розрахунок за точною формулою біноміального розподілу можна замінити наближеним виразом розподілу Пуассона . Відповідь на це питання дає наступна теорема:

$$\left| \sum_k \left(P_k(n) - \frac{\lambda^k}{k!} e^{-\lambda} \right) \right| \le \min(p, np^2). \qquad (6.17)$$

6.3.3. Нормальний розподіл Гауса

Іншим широко розповсюдженим розподілом є нормальний розподіл Гауса. Знайдена Гаусом функція розподілу неперервної випадкової величини X має вигляд:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mathrm{M})^2}{2\sigma^2}}, \qquad (6.18)$$

де M і σ – параметри розподілу.

Параметр M нормального гаусового розподілу дорівнює математичному очікуванню, а параметр σ – середньоквадратичному відхиленню випадкової величини.

Параметри σ і M пов'язані між собою наступним чином:

$$\sigma = \frac{1}{\sqrt{2\pi} f(M)}. \qquad (6.19)$$

Графік функції Гауса для трьох різних значень σ наведено на рис. 6.2. Функція $f(x)$ симетрична відносно ординати, проведеної в точці $x = M$; має максимум при $x = M$ та точку перегину при $x = \pm\sigma$. Таким чином, дисперсія характеризує ширину функції розподілу, або, іншими словами, показує, в якому інтервалі знаходяться значення випадкової величини відносно її середнього значення.

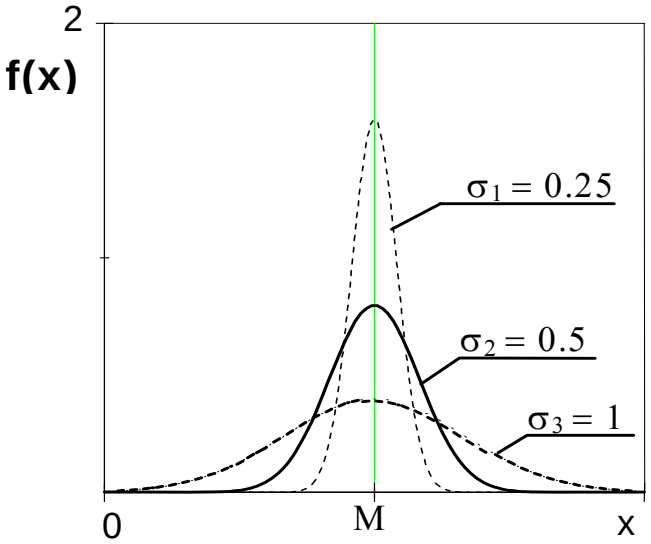


Рис. 6.2

Запитання до розділу 6

1. Що таке дискретні та незалежні випадкові величини?
2. Що таке незалежні події?
3. Чому дорівнює ймовірність достовірної події?
4. Чому дорівнює ймовірність неможливої події?
5. Що таке функція розподілу?

6. Що таке густина розподілу ймовірності?
7. Напишіть закон розподілу ймовірності при незалежних випробуваннях Бернуллі.
8. За яких умов замість закону розподілу ймовірностей Бернуллі можна використовувати закон розподілу Пуассона?
9. Напишіть закон розподілу ймовірностей Пуассона.
10. Які співпадіння вважаються значущими?

Література до розділу 6

1. Дурбин Р., Эдди Ш., Крог А., Митчисон Г. Анализ биологических последовательностей. Вероятностные модели белков и нуклеиновых кислот. – Москва: Ижевск, 2006.
2. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистические методы в медико-биологических исследованиях с использованием Exel. – Киев: Моритон, 2000.

Словник термінів

Escherichia coli – бактерії групи кишкових паличок (сімейство Enterobacteriaceae, рід Escherichia (ешерихії)) - короткі (1-3 мкм) поліморфні рухливі і нерухливі палички, грамнегативні, не утворюють спор, що використовуються в санітарній мікробіології як маркер фекальної контамінації, тобто належать до групи т.з. санітарно-показових мікроорганізмів, оскільки входять в склад нормальної мікрофлори шлунково-кишкового тракту людини, виробляють ряд необхідних для людини вітамінів: B_1, B_2, B_3, B_5, B_6, B_9, B_{12}, К; беруть участь в обміні холестерину, білірубіну, холіну, жовчних і жирних кислот, впливають на всмоктування заліза і кальцію.

Аденін – 6-аминопурин; одна з чотирьох букв генетичного коду - найпоширеніша в природі пуринова основа, що входить до складу нуклеїнових кислот (ДНК та РНК), а також аденозіну, аденозінфосфорних кислот, деяких ферментів.

Алелі або **алельні гени -** різні форми одного і того самого гена, що зустрічаються в межах однієї популяції організмів, розташовані в однакових ділянках (локусах) гомологічних хромосом та визначають альтернативні варіанти розвитку однієї ознаки, тобто різні фенотипи цих організмів.

Алфавіт – літери, що прийняті для опису послідовностей і позначають 4 основи в нуклеотидних послідовностях ДНК або РНК і 20 амінокислот у послідовностях білків.

Амінокислоти – органічні сполуки, в молекулі яких одночасно містяться карбоксильні й амінні групи (амінокарбо́нові кисло́ти), і які є елементарними структурними одиницями білків.

Білки - складні високомолекулярні природні органічні речовини, що складаються з амінокислот, з’єднаних пептидними зв'язками, тобто -COOH (карбоксильна) група однієї амінокислоти реагує з -NH_2 (аміногрупою) іншої.

Біосинтез – 1. Біохімічний процес утворення органічних речовин в живих організмах або поза них з більш простих сполук за участі біокаталізаторів ферментів у процесі обміну речовин. 2. Промислове одержання за допомогою (мікро-)організмів антибіотиків, гормонів, вітамінів, амінокислот та інших біологічно активних речовин.

Ген – структурна і функціональна одиниця спадковості, що контролює зберігання і передачу нащадкам визначеної ознаки або властивості, поняття гену не обмежується ділянкою ДНК, що кодує, але включає й регуляторні послідовності, які можуть знаходитись на відстані мільйонів пар нуклеотидів.

Геном – сукупність всіх генів гаплоїдного набору хромосом і кожного з позахромосомних генетичних елементів, що міститься в окремій клітині зародкової лінії багатоклітинного організму.

Гомеобокс - ділянка ДНК довжиною 183 нуклеотидних залишків, яка кодує гомеодомен.

Гомеодомен – це структурний домен білків, відносно консервативна ділянка білка довжиною в 61 амінокислотний залишок, що зв’язує ДНК або РНК, широко розповсюджений серед факторів транскрипції. Гомеодомен утворює структуру спіраль-поворот-спіраль, в якій альфа-спіралі зв’язані короткими петлевими ділянками. У еукаріот гомеодомени індукують диференціацію клітин, запускаючи каскади генів, необхідних для утворення тканин і органів.

Гуанін – 2-аміно-6-оксипурин; одна з чотирьох букв генетичного коду - найпоширеніша в природі пуринова основа, що є складовою нуклеїнових кислот (ДНК та РНК) або знаходиться в клітині у вільному стані.

ДНК (дезоксирибонуклеїнова кислота) – один з двох типів нуклеїнових кислот, які є основним компонентом клітини, що забезпечує зберігання, передачу і реалізацію генетичної програми розвитку і функціонування живих організмів і до складу якої входять залишок фосфорної кислоти, дезоксирибоза й азотисті основи — аденін, цитозин, гуанін і тимін.

Екзони – ділянки ДНК, які містять інформацію про будову білка і після транскрипції і сплайсингу входять у склад відповідної зрілої мРНК. Екзони більшості генів еукаріотів розділені сегментами некодуючої ДНК (інтронами), які видаляються під час сплайсингу.

Експресія генів – процес, в ході якого спадкова інформація генів перетворюється у функціональний продукт - білок або РНК. Експресія генів може регулюватися на всіх стадіях процесу: і під час транскрипції, і під час трансляції, і на стадії пост-трансляційних модифікацій білків.

Ензими (ферменти) – біокаталізатори білкової або РНК-природи (рибозими). Ферменти специфічно каталізують майже всі хімічні реакції, які відбуваються в живих організмах, мають одну або декілька субодиниць, один або більше активних центрів, сайти зв’язування з інгібіторами або можуть мати кофактор для здійснення каталізу.

Еукаріоти – одно- і багатоклітинні рослинні і тваринні організми (крім бактерій і вірусів), у яких на відміну від прокаріот в клітинах є диференційоване і відділене мембраною ядро, в якому містяться хромосоми.

Імуноглобуліни – антитіла, білкові сполуки, які організм виробляє у відповідь на потрапляння в організм антигенів (чужорідні агенти).

Інсулін – гормон підшлункової залози, що, в першу чергу, регулює вуглеводний обмін в організмі; застосовується для лікування діабету.

Інтрони – ділянки ДНК, в яких на відміну від екзонів, не міститься інформації про послідовність амінокислот білку. Майже всі еукаріотні ядерні інтрони починаються з GU і закінчуються AG (правило AG-GU).

Кодон – одиниця генетичного коду, трійка нуклеотидних залишків в ДНК або РНК, що кодують включення одній амінокислоти.

Ліпопротеїни (ліпопротеїди) – клас складних білків, простетична група яких представлена яким-небудь ліпідом. Так, у складі ліпопротеїдів можуть бути жирні кислоти, нейтральні жири, фосфоліпіди, холестериди.

Мутації – загальна властивість живих організмів, що лежить в основі еволюції і селекції всіх форм життя і являє собою спонтанну зміну в генетичній інформації (заміна, випадання або вставка нуклеотидів в ДНК, РНК.

Нуклеотиди – фосфорні ефіри нуклеозидів, нуклеозидфосфати. Нуклеотиди є складовими частинами нуклеїнових кислот і багатьох коферментів. Вільні нуклеотиди, зокрема АТФ, цАМФ, АДФ відіграють важливу роль в енергетичних і інформаційних внутрішньоклітинних процесах.

Онкоген – це ген, експресія якого призводить до неконтрольованої (злоякісної) проліферації (трансформації) клітин, що може викликати утворення злоякісної пухлини. В онкогени можуть перетворюватись протоонкогени в результаті мутацій, понаднормативної експресії та деяких інших механізмів

Оперон - функціональна одиниці геному у прокаріот, у склад якої входять гени, що кодують білки, що працюють спільно або послідовно (як правило, ферменти, які каналізують окремі етапи одного метаболічного шляху) і об’єднані під одним або декількома промоторами.

Плазміди – позахромосомні фактори спадковості, що являють собою кільцеві (замкнуті) або лінійні молекули ДНК, здатні до автономної реплікації. Використовуються в генно-інженерних маніпуляціях.

Промотор - послідовність нуклеотидів ДНК, впізнавана РНК-полімеразою як старт для початку специфічної транскрипції. Є регуляторною ділянкою ДНК і розташована перед відкритою рамкою зчитування гену або на початку оперону у прокаріот (у напрямку до 5' кінця молекули), забезпечуючи контроль транскрипції цього гену або оперону. Кожний ген або оперон може мати більш ніж один промотор, що регулюються окремо один від одного, та, крім того, інші регуляторні ділянки.

Прокаріоти – одноклітинні організми, що не мають (на відміну від еукаріот) оформленого клітинного ядра і, відповідно, ядерної оболонки; до прокаріотів відносяться бактерії та археї.

Процесінг – 1 Одна із стадій біосинтезу білка між транскрипцією і трансляцією, в якій відбувається видалення інтронів із іРНК, редагування і об'єднання нуклеотидів у зрілу молекулу мРНК. 2 Посттрансляційні модифікації зимогену в активну форму ензиму.

РНК (рибонуклеїнова кислота) - один з двох типів нуклеїнових кислот, які є основним компонентом клітини, що бере участь у реалізації генетичної інформації і біосинтезі білків і до складу якої входять залишок фосфорної кислоти, рибоза і азотисті основи - аденін, цитозин, гуанін і урацил. За будовою і метаболізмом РНК відрізняються функціонально і поділяються на іРНК, тРНК і мРНК.

Секвенування - визначення первинної структури (тобто послідовності мономерів в гетерополімері) нерозгалужених біополімерів: розшифрування

нуклеотидних послідовностей ДНК чи РНК або амінокислот в білку. Сучасні методи секвенування геномів – метод хімічної деградації (Максама-Гілберта) і метод синтезу ДНК на матриці в присутності термінаторів синтезу (за Сенджером).

Сплайсинг – (від англ. splice - з'єднувати), видалення з молекули РНК інтронів і з'єднання ділянок, що залишились і несуть генетичну інформацію (екзонів), в одну молекулу.

Таксономічна група – група організмів, що об'єднані певним ступенем спорідненості, спільними рисами будови та функціональних особливостей. Таксономічними групами можуть бути види або групи видів, популяції, окремі організми або гени.

Тимін - (5-метилурацил); одна з чотирьох букв генетичного коду, найпоширеніша в природі пиримидинова основа, що є складовою ДНК і транспортної РНК.

Транскрипція – процес синтезу іРНК з використанням ДНК як матриці, що відбувається у клітинах, тобто побудова РНК, комплементарної до ДНК.

Тромбоцити – формені елементи крові, що утворюються в кістковому мозку і беруть участь у забезпеченні гемостазу: дрібні (2-4 мкм) пласкі безбарвні без'ядерні клітини неправильної округлої форми.

Урацил - (2,4-диоксопиримидин); одна з чотирьох букв генетичного коду, найпоширеніша в природі пиримидинова основа, що є складовою РНК.

Філогенетичне дерево – дерево, що відображає еволюційні взаємозв'язки між різними видами, іншими таксонами, генами або іншими сутностями, що

мають загального предка. Вершини дерева позначають таксономічні групи, а ребра - відносини спорідненості, як правило, предок - потомок.

Хромосоми – ниткоподібні органоїди еукаріотичного ядра, сукупність яких визначає основні спадкові властивості клітин і організмів; в хромосомах міститься генетична інформація у вигляді генів, що становить близько 90 % ДНК клітини.

Цитозин – 2-окси-6-амінопиримидин, одна з 4 букв генетичного коду, найпоширеніша в природі пиримидинова основа, що є складовою ДНК і РНК.