Збірник наукових праць
VIII Всеукраїнської науково-практичної конференції з міжнародною участю
«Біологічні дослідження – 2017»

УДК 519.257

КОМП'ЮТЕРНІ МЕТОДИ ДОСЛІДЖЕННЯ ЗАКОНОМІРНОСТЕЙ ЗАПИСУ ГЕНЕТИЧНОЇ ІНФОРМАЦІЇ В ДНК

Ю.Г. Вагіс
Національний медичний університет імені О.О. Богомольця, бул. Т. Шевченка, б. 13, м. Київ, 01004, Україна

Двадцять перше століття можна сміливо назвати століттям біології та інформатики. В останні десятиріччя спостерігається бурхливий розвиток досліджень у молекулярній біології та генетиці, який став можливим завдяки використанню математичних методів і комп'ютерних технологій в обробці експериментальних даних і їх аналізі.

Сучасна біоінформатика виникла наприкінці 70-років XX ст. з появою ефективних методів розшифрування послідовностей ДНК. Важливою віхою у становленні та розвитку біоінформатики став міжнародний проект „Human Genome Project", з яким зв'язували великі сподівання вчені з різних галузей науки. Суттєву роль у розвитку технології аналізу генетичної інформації відіграє розвиток комп'ютерної техніки та обчислювальних методів. Відкритість біологічних баз даних, завдяки Інтернет технологіям, дає можливість дослідникам із усього світу та різних галузей науки долучитись до вирішення найсучасніших наукових проблем.

Сьогодні вчені знаходяться на початковому етапі досліджень генетичної інформації про живу матерію, проте розвиток все більш ефективних методів аналізу ДНК та білків, а також розпізнавання біологічних послідовностей дозволяє сподіватися на суттєвий прогрес у розумінні будови та механізмів функціонування живих систем. Створення математичних моделей опису послідовностей ДНК викликають великий інтерес. При вивченні геномів та білків основним завданням є визначення їх структурних, а далі і функціональних особливостей.

Біологічна та медична кібернетика як наукові напрями почали формуватися в Україні на початку 60-х років XX ст. Їх засновниками були видатні українські вчені В.М. Глушков і М.М. Амосов. В Інституті кібернетики АН УРСР було створено відділ біологічної кібернетики. Коло інтересів учених охоплювало не лише медико-біологічні проблеми, а й проблеми пізнання людини та всесвіту в цілому. Саме загальносистемний підхід до вивчення природи людини знайшов відображення у наукових напрямах, ініційованих М.М. Амосовим в галузі кібернетики: моделювання фізіологічних функцій організму людини (фізіологічна біокібернетика), моделювання розумових і психічних функцій людини (психологічна біокібернетика). Він бачив у математичному моделюванні великі можливості дослідження різноманітних явищ природи, в тому числі і при дослідженні організму людини. Була сформована українська школа біологічної та медичної кібернетики й інформатики, яка набула широкого визнання не лише в Україні, а й за її межами [1].

Починаючи з 2003 р. в Інституті кібернетики імені В.М. Глушкова НАН України був проведений статистичний аналіз геномів людини, інших вищих організмів, рослин, бактерій, вірусів з метою виявлення закономірностей у запису генетичної інформації на рівні ДНК хромосом [1, 2]. Перший значний результат – виявлення та математичне обгрунтування властивості симетрії у ДНК. На основі отриманих статистичних даних було показано, що існують строгі правила формування структури ДНК, які справедливі для всіх видів організмів. Досліджено фундаментальні структурні властивості симетрії у запису генетичної інформації в ДНК. Статистичний аналіз підтвердив виконання співвідношень симетрії у геномах бактерій, рослин, вищих організмів, людини [3].

Комп'ютерні підрахунки показали, що в одному ланцюгу ДНК знаходиться приблизно рівна кількість комплементарних нуклеотидів (C – G, A – T): $n(A) \approx n(T)$, $n(C) \approx n(G)$.Тоді, по правилу зв'язку двох ланцюгів ДНК – правилу компліментарності, можна зробити висновок, що має місце симетрія відносно запису нуклеотидів по кожному ланцюгу ДНК [3]: $n(A,1) \approx n(A,2)$, $n(T,1) \approx n(T,2)$, $n(C,1) \approx n(C,2)$, $n(G,1) \approx n(G,2)$ де (*j*,1) – нитка 1, (*j*,2) – нитка 2, $j \in \{A,C,G,T\}$. Знак рівності не ставиться тому, що існує відповідна точність секвенування геномів і вона не відповідає 100% .

Було доведено, що симетрія виконується і для більш довгих послідовностей нуклеотидів, наприклад для пар, тобто по одному ланцюгу кількість пар AC повинна співпадти з кількістю пар GT: *n(AC) ≈ n(GT),* а *n(AG) ≈ n(CT)* тощо.Очевидно, що кількість коротких послідовностей нуклеотидів приблизно збігається з кількістю обернено комплементарних послідовностей нуклеотидів. Для пар нуклеотидів виконуються співвідношення: $n(ij) = n(\bar{j}\bar{i})$, де n-кількість пар, $i, j \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, $\overline{\mathrm{A}} = \mathrm{T}$, $\overline{\mathrm{T}} = \mathrm{A}$, $\overline{\mathrm{C}} = \mathrm{G}$, $\overline{\mathrm{G}} = \mathrm{C}$. Звідси випливає симетрія для пар нуклеотидів щодо двох ланцюгів ДНК: $n(ij,1) = n(ij,2)$.

Кодони (трійки) нуклеотидів зв'язані по одному ланцюгу наступними співвідношеннями: $n(ijk) = n(\bar{k}\bar{j}\bar{i})$, де $i, j, k \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, $n(ijk)$ – кількість кодонів (i, j, k), $(\bar{k}\bar{j}\bar{i})$ – кількість антикодонів (ijk). Симетрія для трійок по двох ланцюгах має вигляд: $n(ijk,1) = n(ijk,2)$.

Комп'ютерні обчислення проводилися для різних геномів і показали, що співвідношення симетрії виконуються також для більш довгих послідовностей нуклеотидів – до десяти літер.

Вперше співвідношення симетрії формалізовані у вигляді математичних формул, на відміну від робіт, що проводилися в цьому напрямку за кордоном. Отримані формули значно спрощують сприйняття результатів і є основою для побудови математичного апарата з метою отримання нових висновків щодо структури ДНК.

Було доведено ряд тверджень стосовно властивостей симетрії та отримано правила зниження та підвищення симетрії. Тобто доведено, що із симетрії послідовностей нуклеотидів випливає симетрія коротких послідовностей, аж до одиничних нуклеотидів – зниження симетрії. Наприклад, з симетрії пар нуклеотидів випливає симетрія одиничних нуклеотидів, а з симетрії трійок випливає симетрія пар:

n(A)= n(AA)+n(AC)+n(AG)+n(AT), n(T)=n(TA)+n(TC)+n(TG)+(TT),...;

n(AA)=n(AAA)+n(AAC)+n(AAG)+n(AAT),... .

Використовуючи властивості імовірносної моделі Маркова, показано, що з симетрії пар нуклеотидів випливає симетрія коротких послідовностей нуклеотидів – підвищення симетрії.

Марковські моделі – одні з основних математичних моделей, які використовуються для розпізнавання амінокислотних та нуклеотидних послідовностей. В рамках прихованих марковських моделей (Hidden Markov models (HMM)) передбачається, що послідовність спостережуваних станів прогнозується за допомогою неспостережуваних станів (прихованих) станів. Велику практичну цінність в біоінформатиці має задача пошуку оптимального ланцюжка прихованих станів по відомому спостережуваному ланцюжку. Дійсно, якщо зіставити приховані стани з характеристиками ДНК і білків, які складно отримати експериментально, наприклад просторова структура білка, функціональні ділянки генів, то стає можливим передбачити ці характеристики на основі послідовностей нуклеотидів або амінокислот. В роботі [3] показано як на основі HMM (екзони та інтрони) побудовані прості методи розпізнавання фрагментів генів в яких кожний прихований стан породжує один спостерігаємий символ.

На кожному ланцюзі ДНК розміщено білок-кодуючи ділянки генів довжиною декілька тисяч нуклеотидів, на яких симетрія також виконується. Відсутність симетрії означала б, що оцінки перехідних імовірностей для ланцюгів Маркова, підраховані на генах на двух різних нитках ДНК, не співпадають між собою. Тобто методи розпізнавання на основі моделей Маркова застосовувати було б неможливо. Наявність симетрії, правил зниження та підвищення симетрії дає можливість застосовувати весь потенціал моделей Маркова при вивченні генетичних послідовностей ДНК та білків.

Вагомий аргумент на користь важливості феномену симетрії – дослідження стійкості стандартного генетичного коду при випадкових мутаціях нуклеотидів у кодоні [3].

Література

1. Сергієнко І.В. Наукові ідеї В.М. Глушкова та розвиток актуальних напрямків інформатики – К.: Наук. думка, 2013. – 285 с.

2. Гупал А.М., Сергиенко И.В. Оптимальне процедуры распознавания – К.: Наук. думка, 2008. – 231 с.

3. Гупал А.М., Сергиенко И.В. Симметрия в ДНК. Методы распознавания дискретных последовательностей – К.: Наук. думка, 2016. – 227 с.