Технічні науки

УДК 577

Климчук Марія Ігорівна
студентка
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»
Климчук Мария Игоревна
студентка
Национального технического университета Украины
«Киевский политехнический институт имени Игоря Сикорского»
Klymchuk Mariia
Student of the
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

Кисляк Сергій Володимирович
старший викладач
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Кисляк Сергей Владимирович
старший преподаватель
Национальный технический университет Украины
«Киевский политехнический институт имени Игоря Сикорского»
Kysliak Serhii
Senior Lecturer
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

ПРОГРАМНИЙ ПРОДУКТ ДЛЯ ІДЕНТИФІКАЦІЇ CpG-ОСТРІВЦІВ
ПРОГРАММНЫЙ ПРОДУКТ ДЛЯ ИДЕНТИФИКАЦИИ CpG-ОСТРОВКОВ
SOFTWARE APPLICATION FOR FINDING CPG ISLANDS

Анотація. *В статті розглянутий алгоритм пошуку CpG-острівців за допомогою прихованих Марковських моделей. Створений програмний продукт, що дозволяє ідентифікувати білоккодуючі ділянки геному відповідно до алгоритму декодування Вітербі та візуалізовувати шлях найбільшої ймовірності в біологічних послідовностях еукаріот відповідно до значень перехідних та емісійних ймовірностей.*

Ключові слова: *приховані Марковські моделі, алгоритм декодування Вітербі, CpG-острівці, матриці перехідних та емісійних ймовірностей, епігенетика.*

Аннотация. *В статье рассмотрен алгоритм поиска CpG-островков с помощью скрытых Марковских моделей. Созданный программный продукт, позволяет идентифицировать белоккодирующие участки генома в соответствии с алгоритмом декодирования Витерби и визуализировать путь наибольшей вероятности в биологических последовательностях эукариот в соответствии с значениями переходных и эмиссионных вероятностей.*

Ключевые слова: *скрытые Марковские модели, алгоритм декодирования Витерби, CpG-островки, переходные и эмиссионные вероятности, эпигенетика.*

Summary. *The article discusses an algorithm for searching for CpG islands using hidden Markov models. The created software product makes it possible to identify protein-coding regions of the genome in accordance with the Viterbi*

decoding algorithm and to visualize the path of the highest probability in biological sequences of eukaryotes in accordance with the values of the transition and emission probabilities.

Key words: *Hidden Markov Models, Viterbi decoding algorithm, CpG-islands, transition and emission probabilities, epigenetics.*

Вступ. Розвиток та становлення сучасної біоінформатики обумовлений появою новітніх методів секвенування нуклеотидних послідовностей та розробкою ефективних методів їх асемблювання. Експоненційне накопичення молекулярно-біологічних даних, вимагає від дослідників вирішення однієї з основних проблем біоінформатики, що пов`язана з наявністю невеликої кількості описаних біологічних послідовностей, що зберігаються у базах даних (наприклад Uniprot, Genbank), у порівнянні з тими, що потребують повного анотування. Удосконалення та оптимізація основних алгоритмів та базових методів біоінформатики, а також інтеграція у напрямку математичних наук, можливо, дозволить вирішити основну проблему сучасної біоінформатики. Кожний етап біоінформаційного аналізу, починаючи з секвенування, асемблювання, картування, ідентифікації кодуючих ділянок тощо, пов`язані з застосуванням певних алгоритмів. Враховуючи отримані сучасні знання та стрімкий розвиток такого напрямку генетики, як епігенетика, особливу увагу зі сторони дослідників заслуговують нетривіальні алгоритми, що базуються на прихованих Марковських моделях (далі по тексту HMM). Такий підхід, з урахуванням молекулярної організації генів, дозволяє ідентифікувати білоккодуючі ділянки ДНК еукаріотичних організмів. При цьому знайдені динуклеотиди цитозин та гуанін можуть вказувати на ген, що може бути розташований у напрямку 3′ кінця відповідно до ідентифікованих динуклеотидів.

Мета роботи: розробити програмний продукт для ідентифікації білоккодуючих ділянок генів еукаріот.

Виклад основного матеріалу. ДНК людини складається з чотирьох нуклеотидів, які записуються літерами: A, C, G і T. Дуже цікавими є такі ділянки ДНК, в яких представлена велика кількість нуклеотиду C – цитозин та нуклеотиду G – гуанін. Цитозин, який розміщений за гуаніном є динуклеотидом і називається CpG, де літера «p» - означає фосфатний зв'язок між цими двома нуклеотидами [1, с. 1; 4, с. 1-12]. Цитозин в структурі динуклеотиду CpG може піддаватись метилюванню. Статус метилювання цитозину в динуклеотидах CpG відіграє регулюючу роль в експресії генів. Динуклеотиди CpG часто скупчені в регіонах, які мають назву CpG-острівці або скорочено - CGIs. Дані острівці найчастіше розміщуються у 5′ промоторній ділянці генів, що приймає участь у першому етапі реалізації генетичної інформаціїї - ініціації транскрипції. Якщо цитозин в структурі CpG метильований, 3′ кодуюча ділянка гена не транскрибується – ген знаходиться у «вимкненому» стані. Демитилювання цитозину в структурі CpG острівців, що розташований біля 5′ промоторної ділянки, відновить експресію такого гена. Так званий епігенетичний профіль є об'єктом сучасної науки про вивчення активності генів - епігенетики [28, с. 30-33].

Інформативним для дослідників є координати розташування CpG острівців, що можуть вказувати на нові гени, які у свою чергу можуть відігравати велику роль у появі та розвитку різних генетичних захворювань, включно з онкологічними. Поява C та G нуклеотидів відносно нуклеотидів A та T – аденіну та тиміну, піддається аналізу за допомогою методів машинного навчання, серед яких не останнє місце займають HMM. Крім того, HMM є одним з найпопулярніших способів ідентифікації та прогнозування ділянок концентрації нуклеотидів та динуклеотидів. HMM визначається набором

прихованих станів, кожен з яких має обмежену кількість переходів між іншими станами. Дана модель має початковий та кінцевий стан, та будь-який шлях від початку в кінець буде генерувати певну послідовність.

Найбільш популярними алгоритмами, які можна застосувати для вирішення задачі ідентифікації генів еукаріот за допомогою HMM є: «Forward» та «Backward» алгоритми, а також алгоритм декодування Вітербі.

В даній роботі, було застосовано алгоритм декодування Вітербі, тому розглянемо його детальніше. Алгоритм Вітербі – це алгоритм пошуку найкращого оптимального шляху, який буде мати найбільшу ймовірність. Даний алгоритм належить до алгоритмів методом динамічного програмування. Алгоритм був розроблений американським інженером та бізнесменом Ендрю Джеймсом Вітербі [26, с. 8]. Даний алгоритм відповідає на питання, який шлях буде мати найбільшу ймовірність, враховуючи дані та навчені матриці параметрів (перехідні йомовірності між станами та емісійні йомовірності генерацііі відповідних нуклеотидів у заданому стані)

Алгоритм Вітербі, дозволяє вирішити задачу декодування. Сутність декодування для нашої задачі полягає у тому, що серед усіх станів A,T,G,C, буде знайдений шлях найбільшої вірогідності, що перетинається зі станами CpG, у яких цитозин може метилюватися. Побудуємо модель λ та послідовність спостережуваних станів – E. Визначемо послідовність внутрішніх станів – S^*, які максимізують ймовірність $P(E,S|\lambda)$, (1) [3, с. 49-61]:

$$p^* \equiv P(E,S|\lambda) \equiv \max_S\big(P(E,S|\lambda)\big),$$
$$S^* \equiv \arg\max_S\big(P(E,S|\lambda)\big). \quad (1)$$

Як таких недоліків алгоритм не має, але можна підкреслити, що, по-перше, ми не можемо стовідсотково сказати, що отриманий шлях найбільшої ймовірності буде найкращим, адже можливо були й інші шляхи, які могли бути би кращими. По-друге, алгоритм Вітербі має експоненціальне зростання кількості шляхів, тому що необхідно проаналізувати усі можливі шляхи, при чому більшість обчислень, які виконуються не будуть мати жодного сенсу. Також, при програмній реалізації даного алгоритму, для декодування одного інформаційного символу необхідне виконання великої кількості операцій множення. Дана проблема була вирішена переходом в логарифмічний простір, де операція множення замінюється операцією додавання. Таким чином процедура декодування Вітербі спрощується [27, с. 20-22].

Оскільки в даній роботі, найбільший інтерес полягає у виявлені CpG-островків в геномній послідовності, то прихована Марковська модель повинна визначати два приховані стани:

- фоновий «-», який немає CpG-острівці (A-, C-, G-, T-)
- острівний «+», який має CpG- острівці (A+ C+ G+ T+) (рис. 1).

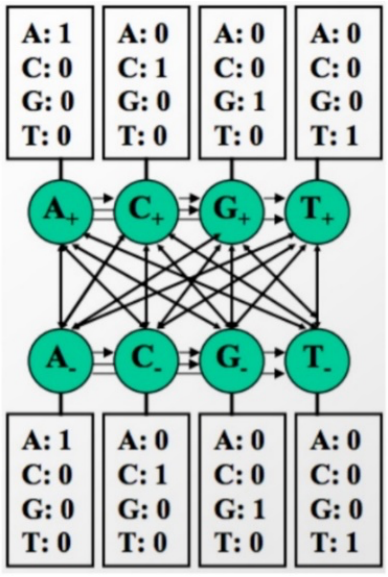

Рис. 1. ПММ для знаходження CpG-острівків. Стрілки вказують на переходи між станами, 0 та 1 – означають емісійні ймовірності [2, с. 46-68]

CpG-острівці є індикатором можливої присутності генів, тому що приблизно 60-70% генів людини містять дані ділянки зі сторони 5′ регуляторної ділянки гену, які маскують промоторні та екзонні ділянки. Динуклеотид CpG називають «гарячою» точкою мутацій, адже з плином часу динуклеотиди в геномі вироджуються. Механізм, який відповідає за дану мутацію – є підвищена вразливість метильованих цитозинів у CpG-острівців до спонтанного дезамінування тиміну (рис. 2). У зв'язку з цим в організмі людини можуть виникати позитивні або негативні зміни.

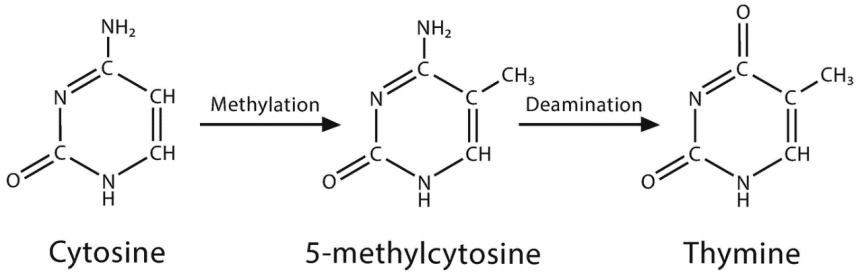

Рис. 2. Схема перетворення цитозину в тимін [5, с. 47-64]

Відомий шотландський біолог та генетик Конрад Х. Ваддінгтон вперше ввів термін «епігенетика» у 1942 році [6, с. 503-518; 7, с. 18-20]. Саме слово «епігенетичний» – означає доповнення до змін у генетичній послідовності, тобто це процес, який змінює активність гена, але при цьому не змінює саму послідовність ДНК. В цьому і є основна різниця між епігенетичними та генетичними механізмами [8, с. 160-167].

Епігенетика є одним з напрямків сучасної генетики, в якій метилювання цитозину CpG-островців відіграють основну роль в процесах регуляції експресії генів на різних етапах онтогенезу. Вивчення взаємозв'язку метилювання з CpG-островками є досить популярним напрямком досліджень. У 2002 році у статті Берда вперше було опубліковано, що дійсно існує зв'язок між CpG-островками та метилюванням [9, с. 6-21]. Крім того, в нещодавніх роботах Роберта С. Іллінгворта та Адріана Берда, було підкреслено, що

метилювання CpG-островків можливе не тільки між різними тканинами, але й між нормальними та злоякісними клітинами, що призводить до подавлення гена, який повинен давати білковий продукт на етапі трансляції [1, с. 10; 10, с. 1713-1720].

Метилювання ДНК – це хімічний процес модифікації ДНК, який може передаватися з покоління в покоління без змін послідовності ДНК. Вперше, процес метилювання ДНК було підтверджено у 1983 році при захворюванні раком, і з того часу його почали спостерігати при багатьох інших захворюваннях [11, с. 21-33; 12, с. 683-692; 6, с. 503-518; 13, с. 37-50]. Даний процес є найбільш вивченим механізмом епігенетичної регуляції генів [14, с. 499-514] та має дуже важливе значення для клітинного перепрограмування, диференціації тканин та нормального розвитку, пов'язаного з багатьма біологічними процесами, включаючи регуляцію експресії генів [13, с. 37-50].

Процес метилювання полягає в додаванні метильної групи – CH_3 до цитозинових основ ДНК, що знаходяться в складі CpG динкулеотида (рис. 3). Метилювання ДНК CpG-островків регулює експресію генів шляхом подавлення транскрипції [16, с. 3157-3173; 17, с. 553-568; 18, с. 142-148], а також має важливе, вирішальне значення для експресії генів та тканино-специфічних процесів. Вважається, що метилювання ДНК властиве еукаріотам, в той час як у людини метилюванню підлягає лише один відсоток геномної ДНК [4, с. 1-12].

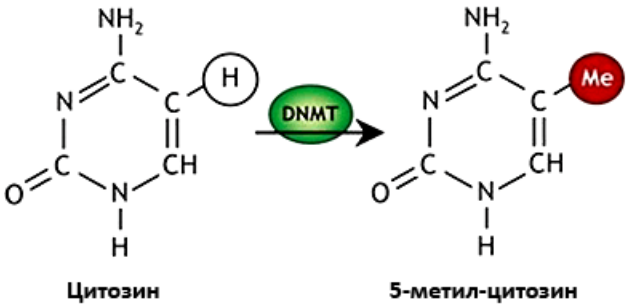

Рис. 3. Схема метилювання цитозину. Зелений овал означає головний фрагмент метилювання, а червоний кружечок означає метильну групу [15, с. 805-832]

З усіх епігенетичних механізмів саме метилювання ДНК викликає найбільший інтерес [20, с. 89-97; 21, с. 5-10], тому що воно бере участь в регуляції генів, а також впливає на транскрипцію генів. З практичної точки зору процес метилювання ДНК має прямий зв'язок з емоційним станом, мозковою діяльністю та навіть з нашим раціоном харчування. В багатьох дослідженнях показано, що зміна метилювання ДНК має великий вплив на ембріональний розвиток, геномний імпринтинг, стабільність генному і статус хроматину [22, с. 597-610]. Також були виявлені зміни в профілі метилювання ДНК у багатьох типах пухлин [23, с. 1156-1163; 24, с. 1-7; 25, с. 357-365].

Тому, метилювання ДНК розглядають як «п'яту основу» геному, яка представляє великий інтерес серед науковців та дослідників.

Створено програмний продукт, який дозволяє ідентифікувати ділянки, в яких можливо буде знаходитися ген, що кодує білок. На рис. 4 показано головну сторінку інтерфейсу програми.

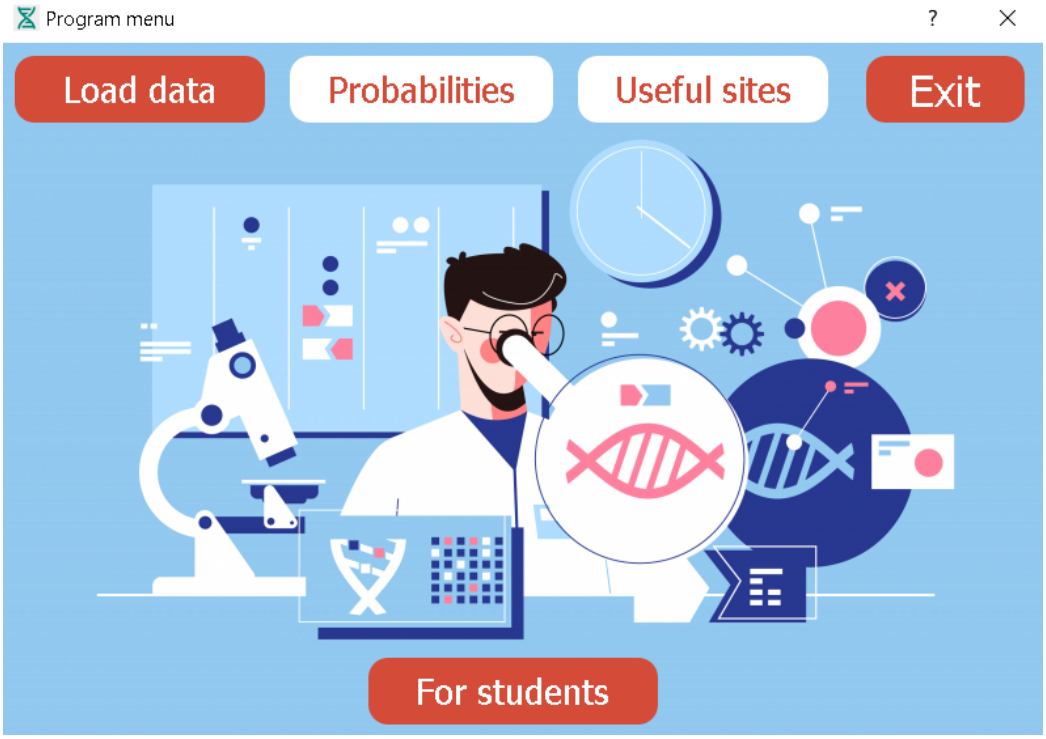

Рис. 4. Вигляд головного вікна інтерфейсу

Для початку роботи з програмою потрібно натиснути на кнопку «Load data», після чого користувач потрапить в нове віконце, де він зможе обрати відповідний файл послідовності та натиснувши на кнопку «Run the algorithm» буде виведено результат роботи програми. Крім того, є можливість зберегти результат до файлу для зручного перегляду усіх результатів (рис. 5). Послідовності нуклеотидів можуть бути завантажені у FASTA форматі.

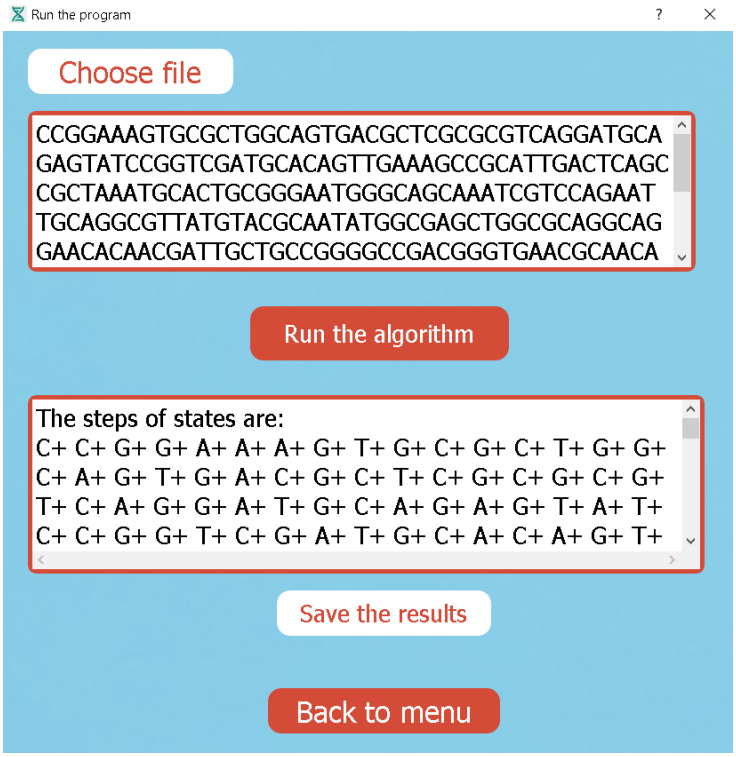

Рис. 5. Демонстрація роботи програми

Також варто розглянути ще один функціонал програми. При натисканні на кнопку «For students» на головному вікні, користувач зможе отримати в результаті не тільки шлях найбільшої ймовірності, але й результуючу матрицю. Дана функція була розроблена для користувачів, які є викладачами вищих навчальних закладів. За допомогою даної програми, вони зможуть наочно в реальному часі продемонструвати роботу алгоритму та відповідно зробити певні висновки щодо декодування послідовності нуклеотидів (рис. 6).

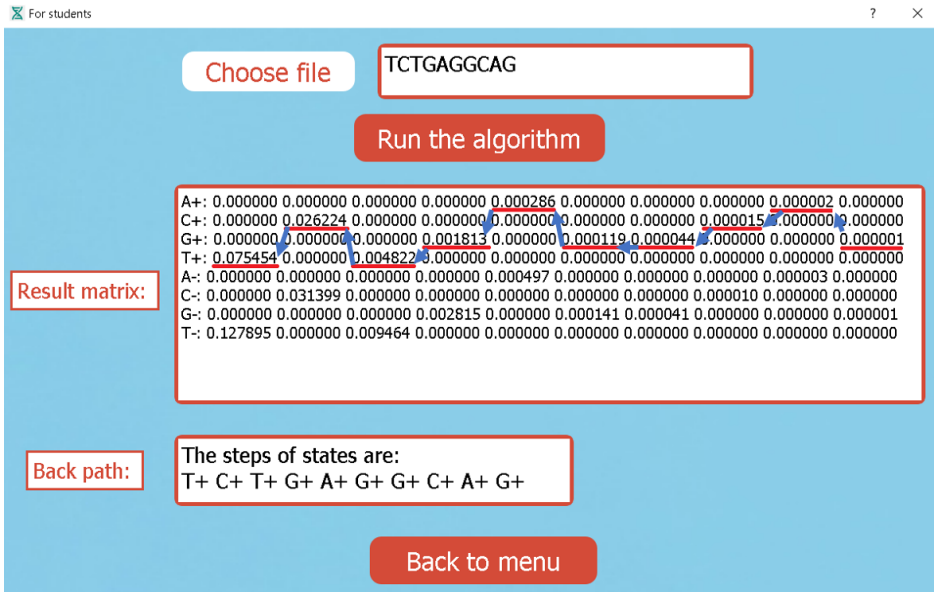

Рис. 6. Вивід результату роботи програми. Червоним підкреслено зворотній шлях найбільшої ймовірності

Крім того, є ще декілька можливостей програми. Користувач може переглянути матриці емісійних та перехідних ймовірностей, які були використані у програмі, натиснувши на кнопку «Probabilities», а також натиснувши на кнопку «Useful sites» на головному екрані, користувач зможе перейти на найбільш популярні сайти, які використовують біоінформатики, а саме: Gen Bank – геномний браузер, Blast, Uniprot, Ensembl (рис. 7).

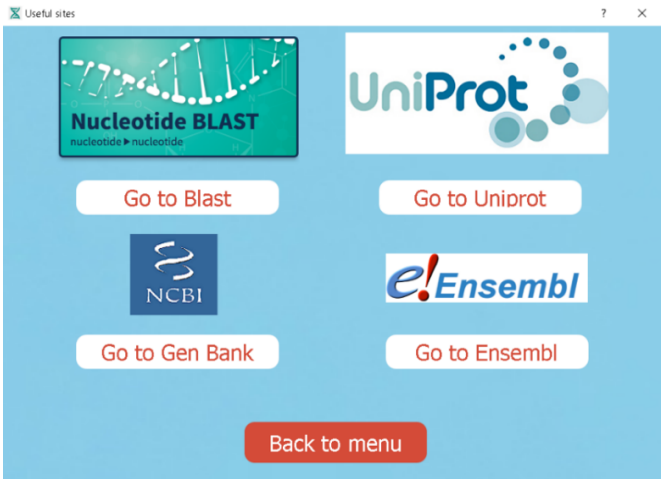

Рис. 7. Вигляд сторінки після натискання на кнопку «Useful sites»

Висновки. Створено програмний продукт зі зручним інтерфейсом, який дозволяє візуалізувати шлях найбільшої ймовірності у вигляді графу з вказаними переходами між станами, а також дозволяє ідентифікувати чи має та чи інша біологічна послідовність кодуючі ділянки у напрямку 3′ кінця ДНК. Користувач може не тільки отримувати результат, але й зберігати його у файл, для подальшого аналізу. Результати роботи можуть бути впроваджені в навчальний процес при викладанні дисципліни «Аналіз біологічних послідовностей». В даній програмі у зручній формі зібрані посилання на найпопулярніші бази даних біологічних послідовностей та сервіси, що можуть бути застосовані для вирішення різноманітних задач біоінформатики.

Література

1. Berg Arnie Exploring the behaviour of the Hidden Markov Model on CpG island prediction. A Thesis Submitted to the College of Graduate Studies and Research in Partial Fulfillment of the Requirements for the degree of Master of Science in the Department of Computer Science University of Saskatchewan Saskatoon. 2013. 87 p.2. Saaty T. L. Analytical planning. the organization of systems / T. L. Saaty, K. P. Kearns. Pergamon Press, 1985. 212 p.
2. Durbin R., Eddy R.S., Krogh A., Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge: Cambridge University Press. 1999. 356 p.
3. Valeria De Fonzo, Filippo Aluffi-Pentini, Valerio Parisi. Hidden Markov Models in Bioinformatics // Current Bioinformatics. 2007. Vol. 2. P. 49-61.
4. Macauley Matthew Identifying CpG islands using hidden Markov models. Department of Mathematical Sciences Clemson University. 2016. P. 1-12.

5. Luke B. Hesson, Antonia L. Pritchard. The DNA Methylation Machinery. Clinical Epigenetics. Springer, Singapore. 2019. P. 47-64.
6. Tabitha M. Hardy, Trygve O. Tollefsbol. Epigenetic diet: impact on the epigenome and cancer // Epigenomics. 2011. Vol. 4. No. 3. P. 503-518.
7. Waddington C.H. The epigenotype. Endeavour. 1942. P. 18–20.
8. Weinhold Bob Epigenetics: The Science of Change // Environ Health Perspect. 2006. Vol. 114. No. 3. P. A160-A167.
9. Берд Адриан Паттерны метилирования ДНК и эпигенетическая память // Гены и развитие. 2002. Vol. 16. No. 1. P. 6-21.
10. Robert S. Illingworth and Adrian P. Bird CpG islands 'a rough guide' // FEBS letters. 2009. Vol. 583. No. 11. P. 1713-1720.
11. Feinberg A.P., Ohlsson R., Henikoff S. The epigenetic progenitor origin of human cancer // Nat Rev Genet. 2006. Vol. 1. No. 7. P. 21-33.
12. Jones P.A., Baylin S.B. The epigenomics of cancer // Cell. 2007. Vol. 4. P. 128. P. 683-692.
13. Yoo C.B., Jones P.A. Epigenetic therapy of cancer: past, present and future // Nat Rev Drug Discov. 2006. Vol. 1. No. 5. P. 37-50.
14. Hao Wu, Brian Caffo, Harris A. Jaffee, Rafael A. Irizarry. Redefining CpG islands using hidden Markov models. Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA / Andrew P. Feinberg // Department of Medicine and Center for Epigenetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. 2010. Vol. 11. No. 3. P. 499-514.
15. Ванюшин Б.Ф. Эпигенетика сегодня и завтра // Вавиловский журнал генетики и селекции. 2013. Vol.17. P. 805-832.

16. Curradi M., Izzo A., Badaracco G., Landsberger N. Molecular mechanisms of gene silencing mediated by DNA methylation. Mol. Cell. Biol. 2002. P. 3157-3173.
17. Ehrlich M., Lacey M. DNA methylation and differentiation: silencing, upregulation and modulation of gene expression // Epigenomics. 2013. P. 553-568.
18. Newell-Price J., Clark A. J., King P. DNA methylation and silencing of gene expression // Trends Endocrinol. Metab. 2000. P. 142-148.
19. Wilkinson M. F. Evidence that DNA methylation engenders dynamic gene regulation. Proc. Natl. Acad. Sci. U.S.A. 2015.
20. Klose R.J., Bird A.P. Genomic DNA methylation: the mark and its mediators // Trends Biochem Sci. 2006. P. 89-97.
21. Cherniuk N.V., Yatsyshyn R.I., Kovalchuk L.Ye., Kaminskyi V.Ya. Modern views on the role of genetic and epigenetic factors in the formation of bronchial asthma. 2019. Vol. 115. No. 2. P. 5-10.
22. Robertson K. D. DNA methylation and human disease. Nat. Rev. Genet. 2005. P. 597-610.
23. De Jager P. L., Srivastava G., Lunnon K., Burgess J., Schalkwyk L. C., Yu L., et al. (2014). Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat. Neurosci. 2014. P. 1156-1163.
24. Esteller M., Herman J. G. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. J. Pathol. 2002. P. 1-7.
25. Sharma P., Kumar J., Garg G., Kumar A., Patowary A., Karthikeyan G., et al. Detection of altered global DNA methylation in coronary artery disease patients // DNA Cell Biol. 2008. P. 357-365.
26. G. David Forney Jr. The Viterbi Algorithm: A Personal History. 2005. P. 122.

27. Башкиров А.В., Остроумов И.В., Свиридова И.В. Основы помехоустойчивого кодирования, основные преимущества и недостатки алгоритмов декодирования. 2007. Р. 20-22.
28. Ржешевский А., Вайсерман А. Эпигенетика: гены и кое-что сверху // Популярная механика. 2015. № 2. С. 30-33.