Методи и софтуерни инструменти за картиране на геноми, базирани на скрити модели на Марков

Пламенка Боровска

Резюме: *Идентифицирането на непознати гени и картографирането на секвенирания геном е основна задача на функционалната геномика. Една от причините за изостаналостта на функционалната геномика в сравнение със структурната е липсата или недостига на надеждни и ефективни методи и софтуерни инструменти за анализ на сурова геномна информация. Повечето популярни генни детектори третират всеки ген изолирано, независимо от другите, което от биологична гледна точка не е акуратно. Направен е прецизиран сравнителен анализ на съвременните методи за предсказване на гени, както и софтуерните инструменти за генни детектори. Особено внимание е насочено към предимствата и ефективността на скритите модели на Марков за картиране на геноми. Представени са експериментални резултати за анализ на генотипа РНК на пандемичния корона вирус SARS-Cov-2 базирани на картиране със скритите модели на Марков.*

Ключови думи: *in silco технологии, картиране на геноми, генни детектори, биоинформатика, скрити модели на Марков*

Methods and Software Tools for Genome Mapping Based on Hidden Markov Models

Plamenka Borovska

Abstract*: Identifying unknown genes and mapping the sequenced genome is a major task of functional genomics. One of the reasons for the backwardness of functional genomics compared to structural is the lack of reliable and effective methods and software tools for analysis of raw genomic data. Most popular gene detectors treat each gene in isolation, independently of the others, which is not biologically accurate. A precise comparative analysis of modern methods for gene prediction, as well as software tools for gene detectors has been performed. Particular attention is paid to the advantages and effectiveness of Markov's hidden models for genome mapping. Experimental results for analysis of the RNA genotype of the pandemic coronavirus SARS-Cov-2 based on mapping with hidden Markov models are presented.*

Key words: *in silco technologies, genome mapping, gene detectors, bioinformatics, Hidden Markov Models*

1. АКТУАЛНОСТ И ЗНАЧИМОСТ НА НАУЧНАТА ПРОБЛЕМАТИКА

Една от важните задачи на функционалната геномика е да идентифицира и картографира гените в нов секвениран геном. Човешкият геном е секвениран през 2000 г., но досега само 10-15% от гените са идентифицирани и функцията на 99% от човешката ДНК остава неизвестна. В по-голяма степен това важи за хиляди напълно известни геноми на други организми. Една от причините за изостаналостта на функционалната геномика в сравнение със структурната е липсата или недостига на надеждни и ефективни методи за анализ на сурова геномна информация.

Идентифицирането на непознати гени и картографирането на секвенирания геном е основна задача на функционалната геномика. Важна задача на научните изследвания в областта

на молекулярната биология е идентифицирането на регулаторни генетични елементи от секвенираните геноми, които ще бъдат използвани за идентифициране и картографиране на неизвестни гени [1].

Целта на настоящата статия е да се анализират съвременните методи и софтуерни инструменти за картиране на геноми като фокусът е върху потенциала на скритите модели на Марков за приложение в компютърната биология.

2. ПРОБЛЕМЪТ ЗА КАРТИРАНЕ НА ГЕНОМА

Първоначалната фаза на геномиката има за цел да определи начален набор от цели геноми, за да се идентифицират всички гени в генома и последователността на протеините, които кодират. Фактът, че много участъци на ДНК в големите геноми е некодиращ, е усложняващо обстоятелство. Трябва да имаме предвид, че некодиращата ДНК включва интрони в гени, регулаторни елементи на гени, множество копия на гени, включително псевдо-гени, междугенни последователности и разпръснати повторения. Що се отнася до кодиращите области, съществуват няколко вида екзони: първоначални кодиращи екзони, вътрешни екзони и терминални екзони.

Анализират се протеин кодиращите участъци на ДНК с цел да се идентифицират началния и крайния кодон на всеки ген в рамките на генома, като се има предвид, че началният start кодон на гена е “ATG”, а stop кодонът има 3 алтернативи – “TAA”, “TAG” or “TGA”

Важните биологични въпроси, на които трябва да се получат отговори, са:

(1) Кои части от ДНК действително правят нещо?

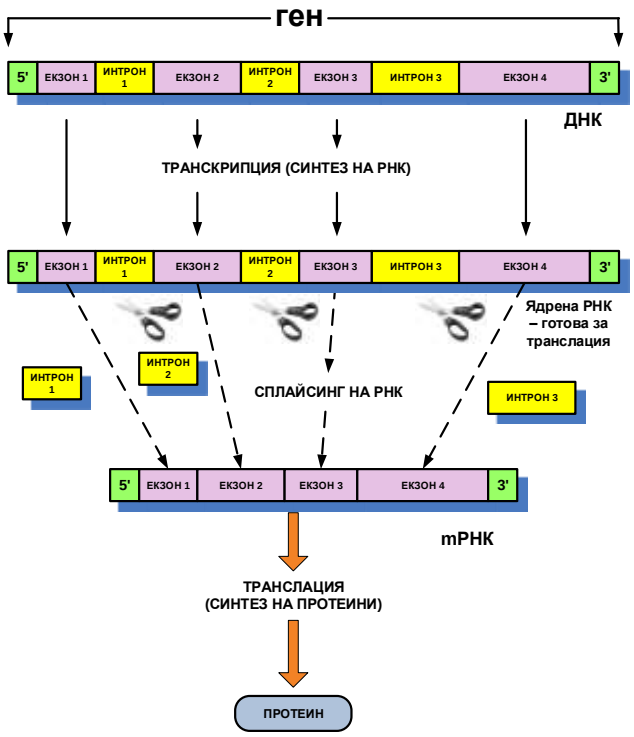
(2) Кои части от ДНК действително кодират протеини или някакъв друг продукт?

(3) Кои части от ДНК регулират експресията?

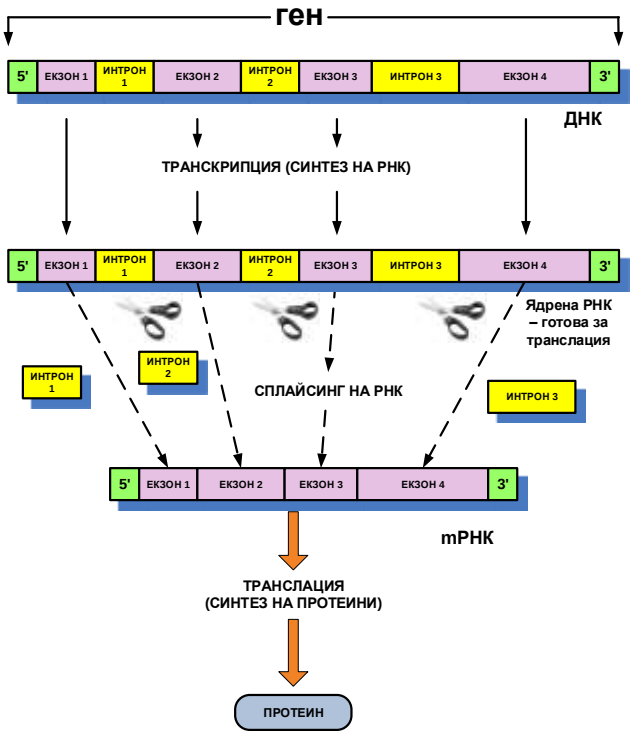
(4) Кои части от ДНК се използват в репликацията?

На фиг. 1 е показан сплайсинг на РНК, а на фиг. 2 – алтернативен сплайсинг на РНК. Различното групиране на екзоните при алтернативния сплайсинг води до синтеза на различен по вид протеини в рибозомата.

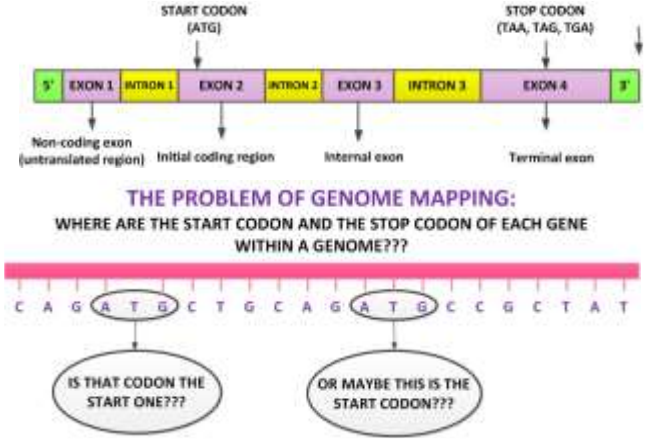
На фиг.1.3 е представен проблемът за картиране на началния кодон на гена.



Фиг.1 Сплайсинг на РНК.



Фиг.2. Алтернативен сплайсинг на РНК.



Фиг.1.3 Проблемът за картиране на началния кодон на гена

3. МЕТОДИ ЗА КАРТИРАНЕ НА ГЕНОМИ

При *емпиричните системи* за откриване на гени (базирани на сходство, хомология или доказателства) се осъществява търсене на таргетния геном в биологичните бази данни при сходни секвенции под формата на известни експресирани секвенции tags, messenger RNA (mRNA), протеинови продукти, и хомоложни или ортоложни секвенции. При зададена секвенция на mRNA, е лесно да се открие уникалната геномна ДНК секвенция от която тя би трябвало да е транскрибирана. При зададена секвенция на протеин може да бъде получена фамилия от възможни кодиращи ДНК секвенции чрез обратна транслация на генетичния код. След като веднъж бъдат определени ДНК секвенциите – кандидати, имаме относително прост алгоритмичен проблем да се търси ефективно таргетен геном за откриване на съвпадения (пълни или частични), точни или приблизителни. Прилагат за алгоритми за локално подравняване, заложени в BLAST, FASTA и алгоритъма на Smith-Waterman за откриване на сходни участъци. Ефективността на този подход се ограничава от съдържанието и акуратността на използваните биологични бази данни.

При прокариотите е изключително важно да се вземе предвид хоризонталния трансфер на гени при търсенето за хомология в секвенцията на гена. Изключително важен фактор както при прокариотите, така и при еукариотите, който не се отразява в съвременните софтуерни инструменти за откриване на гени е съществуването на клъстери от гени (gene clusters) — т. нар. оперони (operons), които представляват функциониращи единици от ДНК съдържащи клъстер от гени под управлението на един и същ промотер.

Повечето популярни генни детектори третират всеки ген изолирано, независимо от другите, което от биологична гледна точка не е акуратно. *В генетиката промоторът е секвенция от ДНК, към която се свързват протеини, инициращи транскрипция на една РНК от ДНК по посока downstream) Тази РНК може да кодира протеин или може да има функция сама по себе си, като tRNA, mRNA или rRNA.* Ribosomal ribonucleic acid (rRNA) е тип некодираща РНК, която е основен компонент на рибозомите и е от съществено значение за всички клетки. Промоторите са разположени в близост до началните места на транскрипцията на гени, upstream на ДНК (към 5 'участъка на смисловата верига). Промоторите могат да бъдат с дължина около 100-1000 базови двойки.

Ab Initio (от самото начало) е метод за предсказването на гени, основан на анализ на съдържанието на ген и откриването на сигнали. Поради присъщите разходи и трудности за получаването на външни доказателства за много гени е необходимо да се прибегне до ab initio откриване на гени, при което в ДНК секвенцията систематично се търсят определени признаци за гени кодиращи протеини. Тези признаци могат да бъдат широко категоризирани като сигнали, специфични последователности, които показват наличието на ген наблизо, или съдържание, както и статистически свойства на самата протеин-кодираща последователност. *Находката на Ab initio ген може да бъде по-точно характеризирана като прогнозиране на гена, тъй като обикновено се изискват външни доказателства за категорично установяване, че предполагаемият ген е функционален.*

В геномите на прокариотите, гените имат специфични и относително добре разбрани секвенции на промотерите (сигнали) като Pribnow box и сайтове на свързване с транскриптиционния фактор, които са лесни за систематично идентифициране. В молекулярната биология, транскрипционен фактор (transcription factor - TF) или специфичен свързващ фактор на ДНК (sequence-specific DNA-binding factor) е протеин, който контролира скоростта на транскрипция на генетичната информация от ДНК към messenger

РНК, посредством свързване към специфична ДНК секвенция. Също така, кодирането на последователността за протеин се случва като една непрекъсната отворена рамка за четене (ORF), която обикновено е дълга сто или хиляди базови двойки.

Статистиката на стоп кодоните е такава, че дори намирането на отворена рамка за четене с такава дължина е доста информативен знак. Тъй като 3 от 64 възможни кодона в генетичния код са стоп кодони, така че, очакванията за стоп кодон са приблизително на всеки 20–25 кодона или 60–75 базови двойки в произволна секвенция.

Освен това протеин-кодиращата ДНК има определена периодичност и други статистически свойства, които са лесни за откриване. Тези характеристики правят намирането на гени при прокариотите сравнително лесно, а добре проектираните системи са в състояние да постигнат високи нива на точност.

В зависимост от предназначението си, гените на прокариотите могат да се разделят на две групи: поддържащи гени (house keeping) и специфични гени. *Поддържащите гени* функционират през цялото време при нормални условия и са ангажирани в ежедневните метаболитни дейности, отговорни за поддържането на клетката. Повечето от поддържащите гени имат общи структурни особености.

При неблагоприятни промени в околната среда като съществена промяна в температурата, промяна в pH и други екологични особености, като например излагане на токсични химикали, липса на хранителни вещества и всеки друг фактор, който не е благоприятен в околната среда, *специфичните гени* се активират за преодоляването на такива враждебни или неблагоприятни ситуации.

Структурните особености на промотерите на специфичните гени, макар че основно имат общи черти, индивидуално се различават слабо един на друг.

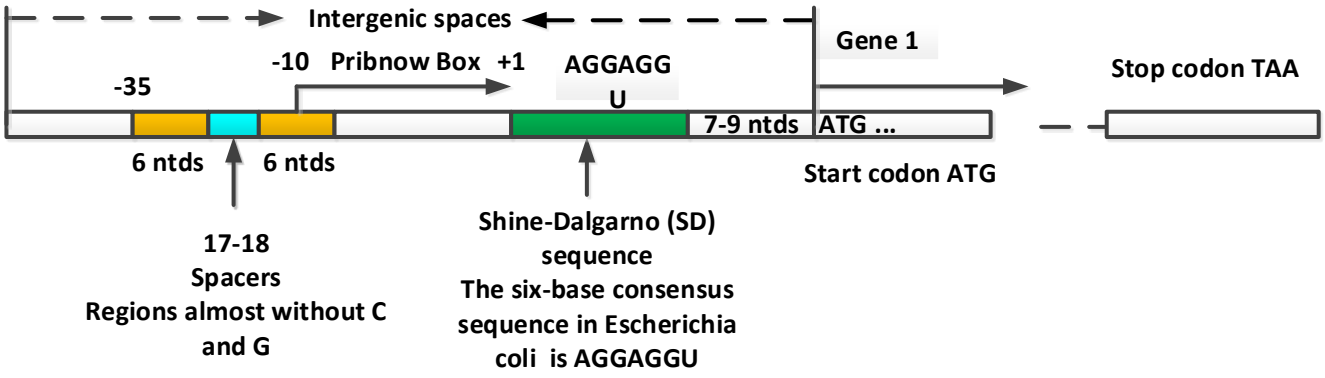
Кодиращият регион започва с инициращ кодон и отворена рамка за четене (ORF) и завършва с терминален кодон. Кодиращият регион на структурните гени не е разделен, но гените на rRNA имат разделители (spacers) помежду си.

Значението на промотера по същество е отделен модул на секвенцията, разпознаван от транскрипционните фактори, които се свързват плътно към секвенцията и инициират транскрипция чрез разтваряне на спирално навитата ДНК в транскрипционен балон. Тези последователности осигуряват информация за сайта, в който ензимът да инициира транскрипцията. Ако някоя от консенсусните области е изтрита или променена значително, ензимът няма да се свърже, или ако се свърже, ще инициира транскрипцията в различни позиции.

Структурата на прокариотен ген е показана на фиг. 1.4.

Намирането на ген Ab initio в еукариоти, особено в сложни организми като хората, е значително по-предизвикателно поради няколко причини. *Първо, промоторът и другите регулаторни сигнали в тези геноми са по-сложни и не толкова добре разбрани, отколкото в прокариотите, което ги прави по-трудни за надеждно разпознаване.*

Два класически примера за сигнали, идентифицирани от гени на еукариоти, са CpG острови и свързващи сайтове за поли (A) опашка.



Фиг.4 Структура на прокариотен ген.

CpG островите (или CG островите) са региони с висока честота на CpG сайтове. Въпреки че обективните дефиниции за островите на CpG са ограничени, обичайната формална дефиниция е регион с поне 200 bp, процент на GC по-голям от 50%, и съотношение на наблюдаван и очакван CpG над 60%.

POLY (A) tail е участък от РНК, който има само аденинови бази. При еукариотите е част от процеса, който произвежда зряла messenger РНК (mRNA) за транслация. Следователно, той представлява част от по-големия процес на генна експресия. Полиаденилирането може да произведе повече от една транскрипция от един ген (алтернативно полиаденилиране), подобно на алтернативния сплайсинг.

Второ, при механизмите за сплайсинг, използвани от еукариотните клетки, определена протеин - кодираща последователност в генома е разделена на няколко части (екзони), разделени от некодиращи последователности (интрони). Софтуерите за откриване на гени при еукариотие често са проектирани да идентифицират сайтовете за сплайсинг. Типичният ген, кодиращ протеина при хората, може да бъде разделен на дузина екзони, всеки с дължина по-малко от двеста базови двойки, а някои са по-малки от двадесет до тридесет. Следователно е много по-трудно да се открият периодичност и други известни свойства на съдържанието на протеинокодиращата ДНК при еукариотите.

4. ПРИЛОЖЕНИЕ НА HMM В КОМПЮТЪРНАТА БИОЛОГИЯ.

Често анализът на биологичната секвенция е просто въпрос на поставяне на правилния етикет върху всяка аминокиселина. *При идентифицирането на гени целта е да се маркират нуклеотидите като екзони, интрони или интергенен участък*. При подравняване на секвенции целта е да се асоциират аминокиселините в секвенцията - заявка с хомоложни секвенции от целеви бази данни. При идентифицирането на гени (gene finder) трябва да бъдат комбинирани следните фактори в единна система за оценяване (scoring system): консенсус на местата на сплайсинг, отклонението на кодоните, предпочитанията за дължината на екзон/интрон, и анализ на отворената рамка за четене.

В теорията на вероятностите моделът на Марков е стохастичен модел, използван за моделиране на случайно променящи се системи. Предполага се, че бъдещите състояния зависят само от текущото състояние, а не от събитията, настъпили преди него (свойството на Марков). Като цяло това предположение дава възможност за разсъждения и изчисления с модела, който в противен случай би бил нерешим. Поради тази причина в областта на прогнозното моделиране и вероятностното прогнозиране е желателно даден модел да прояви свойството на Марков.

Най-простият модел на Марков е веригата Марков (Markov chain). Той моделира състоянието на система с произволна променлива, която се променя във времето. В този контекст свойството на Марков предполага, че разпределението за тази променлива зависи само от разпределението на предишно състояние.

Има четири общи модела на Марков, използвани в различни ситуации, в зависимост от това дали всяко последователно състояние може да бъде наблюдавано или не, и дали системата трябва да се коригира въз основа на направените наблюдения.

	Състоянието на системата е видима	**Състоянието на системата е частично видима**
Автономна система	Верига на Марков (Markov chain))	Скрит модел на Марков (Hidden Markov Model – **HMM**)
Контролирана система	Марков процес на вземане на решение (Markov Decision Process – MDP)	Частично видим процес на решение на Марков

Фиг.5 Общи модели на Марков

Скритите модели на Маркови (HMM) са формална основа за създаване на вероятностни модели на проблеми с „етикетирането“ на линейни секвенции. Те предоставят концептуален инструментариум за изграждане на сложни модели само чрез начертаване на интуитивна картина. Те са в основата на разнообразна гама от програми, включително генериране на данни, търсене на профили, многократно подравняване на секвенции и идентификация на регулаторни сайтове.

HMM е напълно вероятностен модел – всички параметрите на модела и цялостната последователност "резултати" са вероятностни. Следователно, можем да използваме Бейсовата теория на вероятностите, включително за оптимизиране на параметрите и интерпретиране на значимостта на резултатите. Теоремата на Бейс по името на Томас Бейс (*Thomas Bayes*) се използва в теорията на вероятностите за изчисляване на вероятността за настъпване на дадено събитие, след като вече е известна част от информацията за него.

При аналитичен проблем е дадена генетична секвенция и целта е да да направим заключение за пътя на скритото състояние. Има потенциално много пътища за състояния, които биха могли да генерират една и съща секвенция. Основната цел е да се намери пътя на скритото състояние с най-голяма вероятност.

За повечето проблеми има толкова много възможни последователности на състояния, за определянето на които прекомерното голямо изчислително време е неприемливо. Ефективният алгоритъм на Витерби гарантирано намири най-вероятния път на преходите на състоянията, при зададени секвенция и HMM. Алгоритъмът на Витерби е алгоритъм за динамично програмиране, доста подобен на този, използван за стандартно подравняване на секвенции.

"Viterbi path" и "Viterbi algorithm" са се наложили като стандартни термини за приложението на алгоритмите на динамичното програмиране при решаването на MAX проблеми базирани на вероятности. При статистическия анализ (statistical parsing) алгоритъма на динамичното програмиране може да се приложи за контекстно независимо парсиране на стринг, в общия случай известно като парсиране по Витерби ("Viterbi parse“).

Предимство на вероятностното моделиране е, че може да определим директно сигурността (доверителността – confidence) на получените резултати. За широкомащабни проблеми, за целите на пост-декодирането се прилагат два алгоритъма на динамичното програмиране - *Forward и Backward*, които по същество са като алгоритъма на Viterbi, като разликата е, че при тях сумата се формира от сумата на нормализираните вероятности на възможни пътища на състоянията вместо да се избира най-добрият (оптималният) път.

Скритите модели на Марков HMMs не вземат предвид биологичния аспект на корелации между нуклеотидите, тъй като предположението е, че всеки нуклеотид зависи само от едно основно състояние. Пример за този недостатък е, че скритите модели на Марков HMMs обикновено са неподходящи за анализ на вторичната структура на РНК. Консервативните РНК двойки бази при РНК предизвикват широко обхватни корелации между двойките бази; в една позиция може да има произволна база, но партньорът в сдвоената база трябва да бъде комплементарен.

Пътят на състоянията при HMM няма опции за запомняне на резултатите, генерирани от несъседни състояния. В някои случаи е възможно да се променят (нарушат) правилата на HMMs без негативни ефекти върху действието на алгоритмите. За случая на откриването на гени (gene finding), емитирането може да обхваща кодони от тройка корелирани нуклеотида (correlated triplet codon) вместо да се емитират три независими нуклеотида и следователно, алгоритмите базирани на HMM могат да бъдат доразвити за състояния, емитиращи триплети (тройка кодони).

5. СОФТУЕРНИ ИНСТРУМЕНТИ ЗА ОТКРИВАНЕ НА ГЕНИ (*GeneFinders*)

Усъвършенстваните търсачки на гени както за прокариотични, така и за еукариотни геноми обикновено използват сложни вероятностни модели, като например скрити модели на Марков (HMM), за да комбинират информация от различни измервания на сигнали и съдържание. Системата *GLIMMER* е широко използван и с висока точност търсачка на гени за прокариоти. *GeneMark* е друг популярен подход. За сравнение, Eukaryotic ab initio генни търсачки са постигнали само ограничен успех. Забележителни примери са програмите *GENSCAN* и *geneid*.

SNAP gene finder е базиран на HMM като Genscan и се опитва да бъде по-адаптивен към различни организми, като адресира проблеми, свързани с използването на търсач na geni в последователност на генома, срещу която не е бил обучен. Няколко подхода използвани в mSplicer, CONTRAST, или mGene също използват техники за машинно обучение като поддръжка на

векторни машини за успешно прогнозиране на гените. Те изграждат дискриминационен модел, използвайки hidden Markov support vector machines или условни рандомизирани полета, за обучението на акуратна функция за оценка на предсказването на гени. Методите на Ab Initio са тествани с еталонни програми, като някои от тях се приближават до 100% сензитивност. С увеличаването на сензитивността, обаче, точността намалява в резултат на увеличените фалшиви положителни резултати (false positives).

Установено е, че почти всички сплайс сайтове съответстват на консенсусни секвенции. Тези консенсусни секвенции обхващат почти инвариантни динуклеотиди във всеки край на интрона, GT в 5' края на интрона, и AG в 3' края на интрона. Консенсусни секвенции на сплайс сайтовете за U2 (основен клас) интрони в pre-mRNA в общия случай съответстват на:

3' splice sites: CAG|G

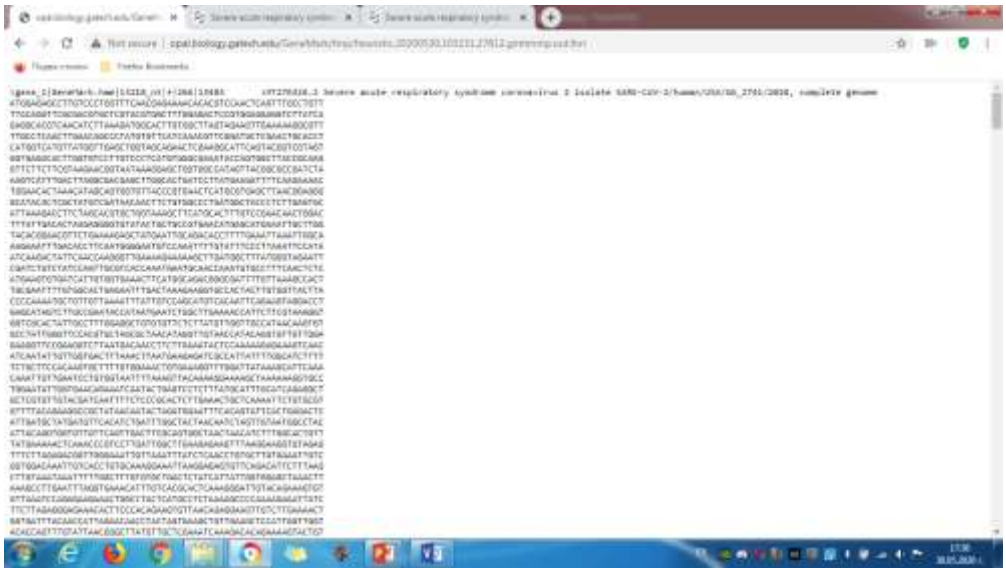
5' splice sites: MAG|GTRAGT където M е A или C, и R е A или G.

GeneSplicer : A computational method for splice site prediction, Johns Hopkins University, Center for Computational Biology, Baltimore, Maryland, представлява бърза и гъвкава система за детектиране на сплайс сайтове в геномна ДНК на различни еукариоти [7].

Genie: Gene Finder Based on Generalized Hidden Markov Models. Създаден е от Computational Biology Group, Университета на Калифорния, Lawrence Berkley National Lab, и Berkley Drosophila Genome Project [8]. Обхваща софтуерни инструменти за анализ за предсказване на промотери (вкл. на основата на изкуствени невронни мрежи), предсказване на сплайс сайт, откриване на гени (gene finder) базирано на скрити модели на Марков [8].

GeneMark е уеб- базиран софтуер за намиране на гени в прокариоти, еукариоти и вируси, за прогнозиране на вирусен ген, и анотиране на вирусни геноми [9]. Генетичните секвенции на вируси, фаги или плазмиди могат да бъдат анализирани или от GeneMark.hmm с евристични модели (ако последователността е по-кратка от 50 kb) или от програмата със самообучение GeneMarkS.

В рамките на това изследване са осъществени експерименти за картиране на пълния геном на пандемичния вирус SARS-CoV-2 от изолати в GenBank (фиг.6).



Фиг. 6. Изолат на пълният геном на пандемичния вирус SARS-CoV-2 в GenBank: MT276326.2

На фиг.8 е представена екранна снимка на резултатите от картирането на пълния геном на пандемичния вирус SARS-CoV-2 от изолат в GenBank LOCUS MT276326 29882 bp RNA linear.

.

5. ЗАКЛЮЧЕНИЕ

В статията е направен прецизиран сравнителен анализ на съвременните методи за предсказване на гени, както и софтуерните инструменти за генни детектори. Особено внимание е насочено към предимствата и ефективността на скритите

модели на Марков за картиране на геноми. Представени са експериментални резултати за анализ на генотипа РНК на пандемичния корона вирус SARS-Cov-2 базирани на карт*иране със скритите модели на Марков.*

GeneMark.hmm PROKARYOTIC (Version 3.26)
Date: Sat May 30 10:18:57 2020
Sequence file name: seq.fna
Model file name: /home/genemark/bin/gmsuite/heu_11.mod
RBS: false

Model information: Heuristic_model_for_genetic_code_11_and_GC_38

FASTA definition line: empty-fasta-def-line
Predicted genes

Gene	Strand #	LeftEnd	RightEnd	Gene Length	Class
1	+	266	13483	13218	1
2	+	13810	21555	7746	1
3	+	21536	25384	3849	1
4	+	25393	26220	828	1
5	+	26523	27191	669	1
6	+	27202	27387	186	1
7	+	27394	27759	366	1
8	+	27894	28259	366	1
9	+	28274	29533	1260	1

Фиг.8. Екранни снимка на резултатите от картирането на пълния геном на пандемичния вирус SARS-CoV-2 от изолати в GenBank със софтуерния инструмент GeneMark - LOCUS MT276326 29882 bp RNA linear

БЛАГОДАРНОСТИ

Резултатите от това изследване са финансирани от научно-изследователски проект ДН07/24/2016г. „Интелигентен метод за адаптивно откриване на in silico знания и взимане на решения, базирани на анализ на потоци големи данни за целите на научните изследвания“, ФНИ, МОН.

REFERENCES

[1] Jin Hwan Do1 and Dong-Kug Choi, Computational Approaches to Gene Prediction, The Journal of Microbiology, April 2006, p.137-144
[2] Dong Yanga, Julian L. Leibowitz, The structure and functions of coronavirus genomic 3_ and 5_ ends, Virus Research 206 (2015) 120–133, www.elsevier.com/locate/virusres
[3] Michael Schatz, Gene Finding & HMMs, Bioinformatics, Quantitative Biology 2014.
[4] Peter Skewes-Cox, Thomas J. Sharpton, Katherine S. Pollard, Joseph L. DeRis, Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data, August 2014 | Volume 9 | Issue 8 | e105067, www.plosone.org
[5] Colin Dewey, Applications of HMMs in Computational Biology, www.biostat.wisc.edu/bmi576.html
[6] Byung-Jun Yoon, Hidden Markov Model and their Application in Biological Sequence Analysis, Current Genomics, 2009, 10, pp. 402-415
[7] GeneSplicer : A computational method for splice site prediction, http://ccb.jhu.edu/software/genesplicer/
[8] Genie: Gene Finder Based on Generalized Hidden Markov Models https://www.fruitfly.org/seq_tools/genie.html
[9] GeneMark, http://opal.biology.gatech.edu/GeneMark/

За автора:

Пламенка Боровска, pborovska@tu-sofia.bg
Професор, доктор на науките
катедра Информатика
Факултет по приложна математика и информатика
Технически университет - София

About the author:
Plamenka Borovska, pborovska@tu-sofia.bg
Professor, DSc, Informatics Dept.,
Faculty for Applied Mathematics and Informatics
Technical University of Sofia