КАРТИРАНЕ НА ГЕНИ

Проф. Боровска

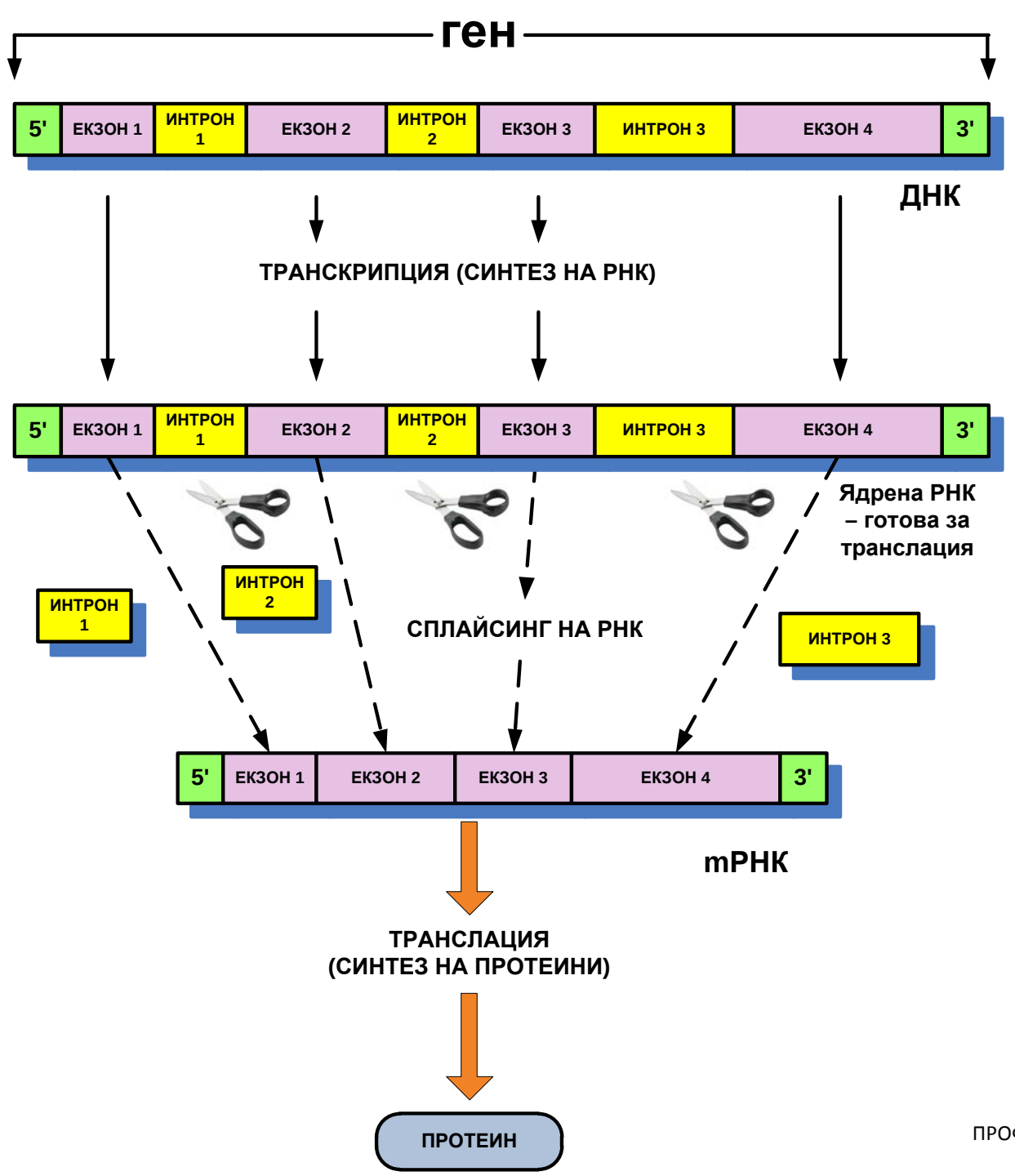
КАРТИРАНЕ НА ГЕНИ

- 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.
- Такива методи могат да бъдат разработени въз основа на генетични регулаторни елементи.
- Те са идеална база за идентифициране и картографиране на гените, тъй като имат характерни структури (нуклеотиден състав и първична структура) и са пряко свързани с функциите на гените.

КАРТИРАНЕ НА ГЕНИ

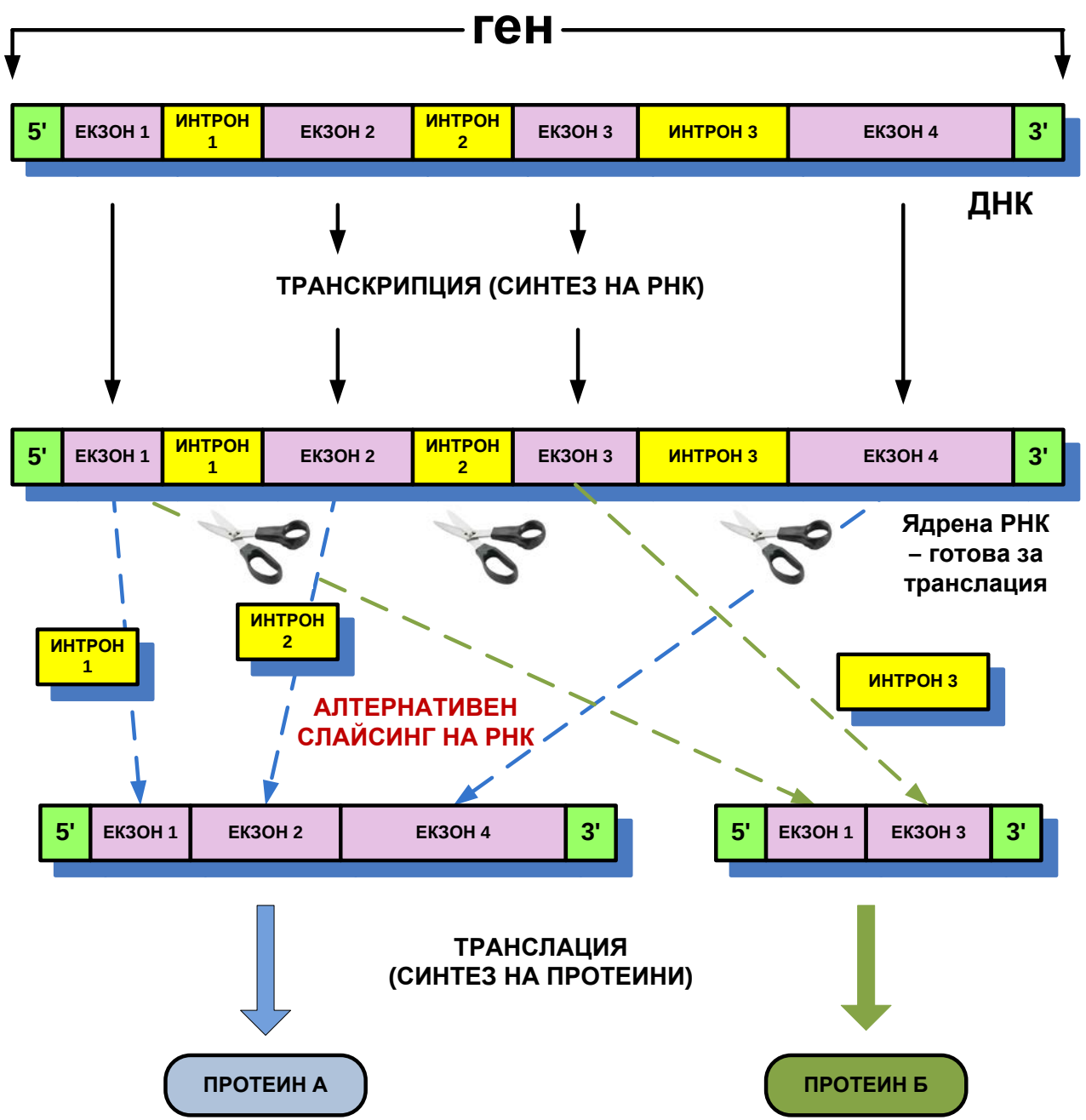
- Идентифицирането на непознати гени и картографирането на секвенирания геном е основна задача на функционалната геномика.
- Важна задача на научните изследвания в областта на молекулярната биология е идентифицирането на регулаторни генетични елементи от секвенираните геноми, които ще бъдат използвани за идентифициране и картографиране на неизвестни гени.
- 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.

СПЛАЙСИНГ НА РНК (RNA SPLICING)



АЛТЕРНАТИВЕН СПЛАЙСИНГ НА РНК (ALTERNATIVE RNA SPLICING)

При алтернативния сплайсинг екзоните се групират по различен начин и в резултат се генерират различни протеини от един и същ ген



ПРОБЛЕМЪТ ЗА КАРТИРАНЕ НА ГЕНИТЕ

- 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”

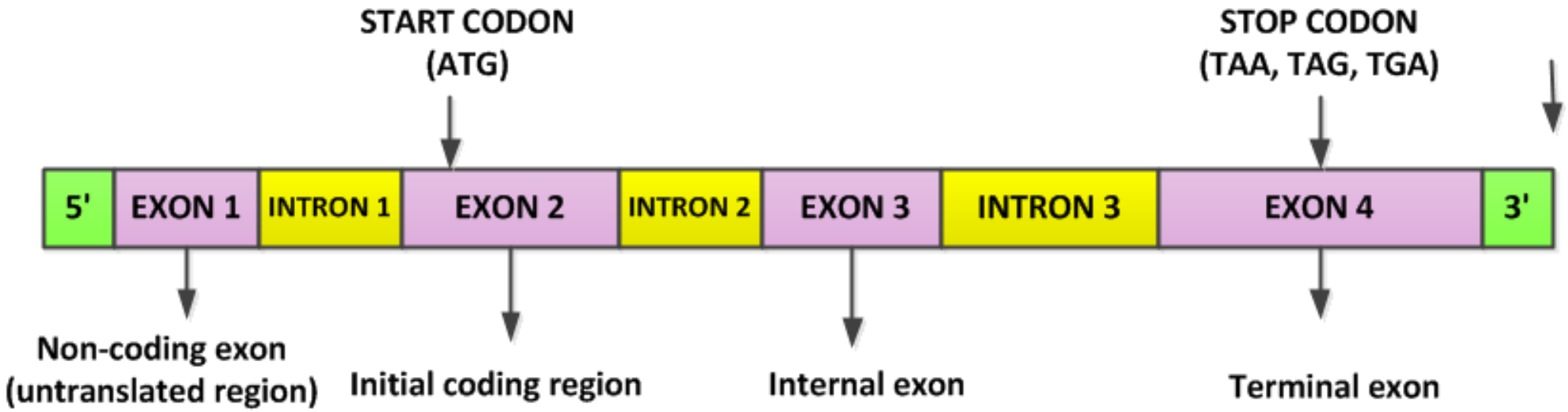
Важните биологични въпроси, на които трябва да се получат отговори, са:

(1) Кои части от ДНК действително правят нещо?

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

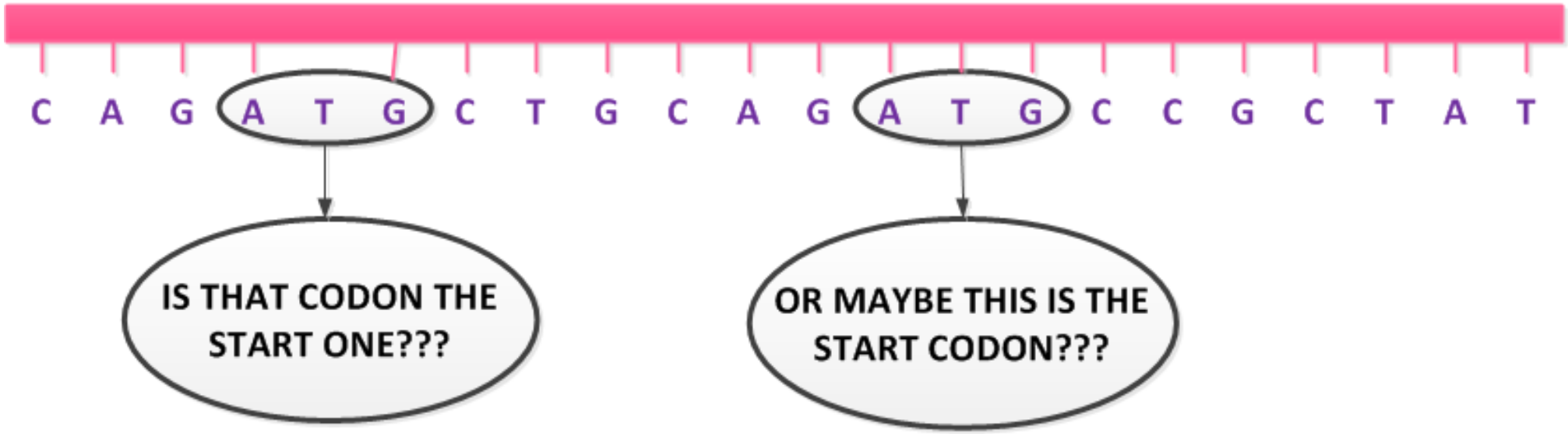
(3) Кои части от ДНК регулират експресията?

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



THE PROBLEM OF GENOME MAPPING:

WHERE ARE THE START CODON AND THE STOP CODON OF EACH GENE WITHIN A GENOME???



ЕМПИРИЧНИ МЕТОДИ

- 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 expressed sequence 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

ПРОМОТЕРИ

- 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, иницииращи 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.

МЕТОДИ AB INITIO

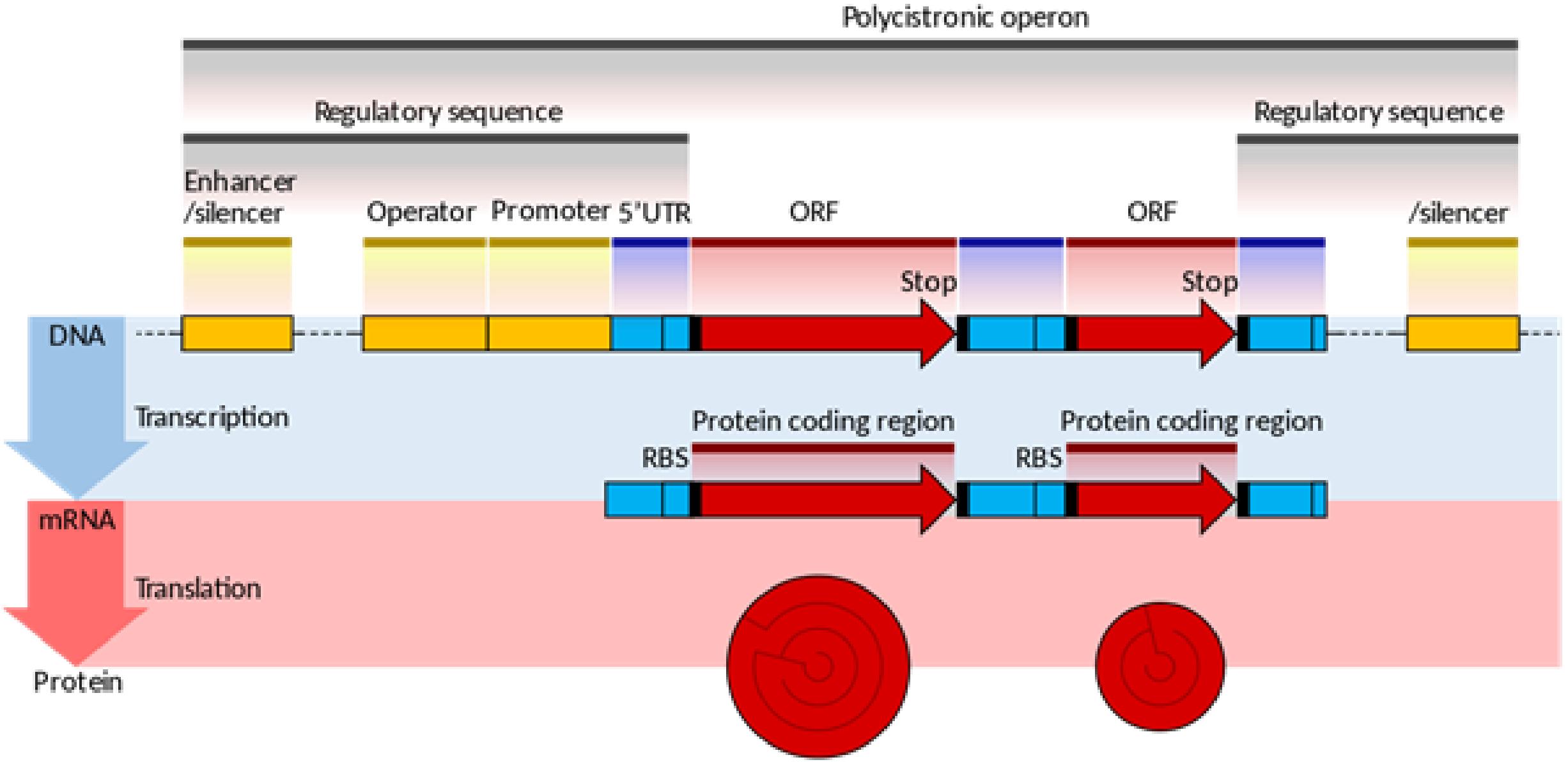
- Ab initio – от самото начало
- Предсказването на гени Ab Initio е метод, основан на анализ на съдържанието на ген и откриването на сигнали.
- Поради присъщите разходи и трудности за получаването на външни доказателства за много гени е необходимо да се прибегне до ab initio откриване на гени, при което 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.*

ПРОКАРИОТИ

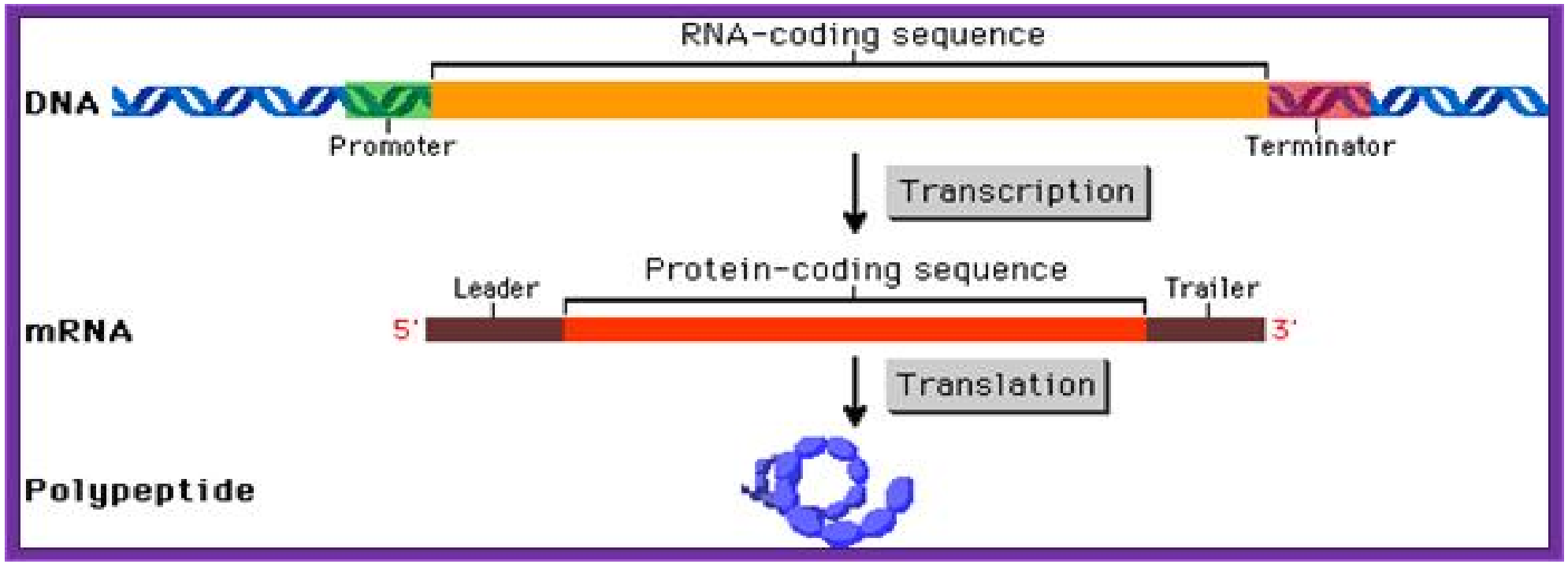
- В геномите на прокариотите, гените имат специфични и относително добре разбрани секвенции на промотерите (сигнали) като Pribnow box и сайтове на свързване с транскриптиционния фактор, които са лесни за систематично идентифициране.
- В молекулярната биология, транскрипционен фактор (transcription factor - 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, човек би очаквал стоп кодон 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.

СТРУКТУРА НА ПРОКАРИОТНИЯ ГЕН

- В зависимост от предназначението си, гените могат да се разделят на две групи: поддържащи гени (house keeping) и специфични гени.
- *house keeping гените* 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 промоторите на специфичните гени, макар че основно имат общи черти, индивидуално се различават слабо един на друг.


Polycistronic operon
Regulatory sequence
Regulatory sequence
Enhancer
/silencer
Operator
Promoter
5'UTR
ORF
ORF
/silencer
Stop
Stop
DNA
Transcription
Protein coding region
Protein coding region
RBS
RBS
mRNA
Translation
Protein

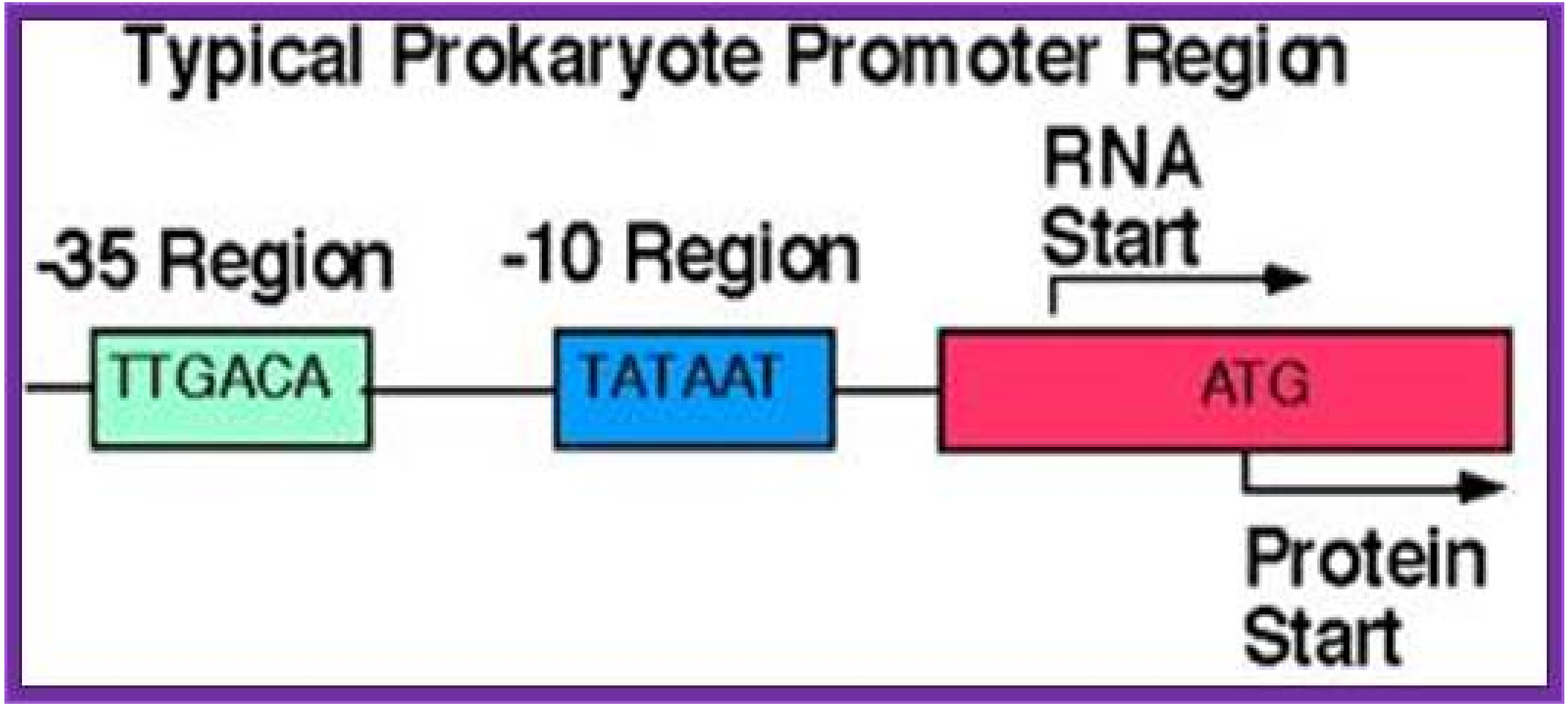
СТРУКТУРА НА ПРОКАРИОТНИЯ ГЕН



- Кодиращият регион започва с иницииращ 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.

СТРУКТУРА НА ПРОКАРИОТНИЯ ГЕН

- Елементите нагоре по веригата (up stream) от началото на кодиращия регион включват промоторни елементи.
- Приблизително на 50 до 100 ntds от старт кодона е първият нуклеотид, от който започва транскрипцията, следователно, в този сайт е първият нуклеотид, който се включва в транскрибираната РНК. Сайтът се нарича *сайт за иницииране на транскрипция или START*.
- Приблизително 10 нуклеотида преди началото, има последователност TATAAT или т. нар. Pribnow box.
- Всеки нуклеотид, намиращ се вляво от началото, се обозначава с (-) символ и регионът се нарича елемент нагоре (upstream element). Номерата са отбелязват като -10, -20, -35 и т.н.
- Стартовият сайт е първият ntd и се отбелязва като +1, като всяка последователност надясно от стартовия сайт се нарича елементи надолу (down stream elements) и се номерират с +10, +35 и т.н.


Typical Prokaryote Promoter Region
RNA
Start
-35 Region
-10 Region
TTGACA
TATAAT
ATG
Protein
Start

ФУНКЦИОНАЛНОСТ НА ПРОМОТЕРА

- Значението на промотора 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.

СТРУКТУРА НА ПРОКАРИОТЕН ГЕН

- При прокариотите един промотер отговаря за последователност от няколко гени
- полицистронна РНК

PRIBNOW BOX

T	A	T	A	A	T
82%	89%	52%	59%	49%	89%

Probability of occurrence of each nucleotide in *E. coli*

The **Pribnow box** (известна също като **Pribnow-Schaller box**) представлява последователност *TATAAT* от 6 нуклеотида, които са съществена част от сайта на промотера за активиране на транскрипцията при бактериите (AT-richness – област богата на AT).

Идеализира се като „*consensus sequence*“— съдържа най-често срещаните бази за всяка позиция при голям брой анализирани промотери;

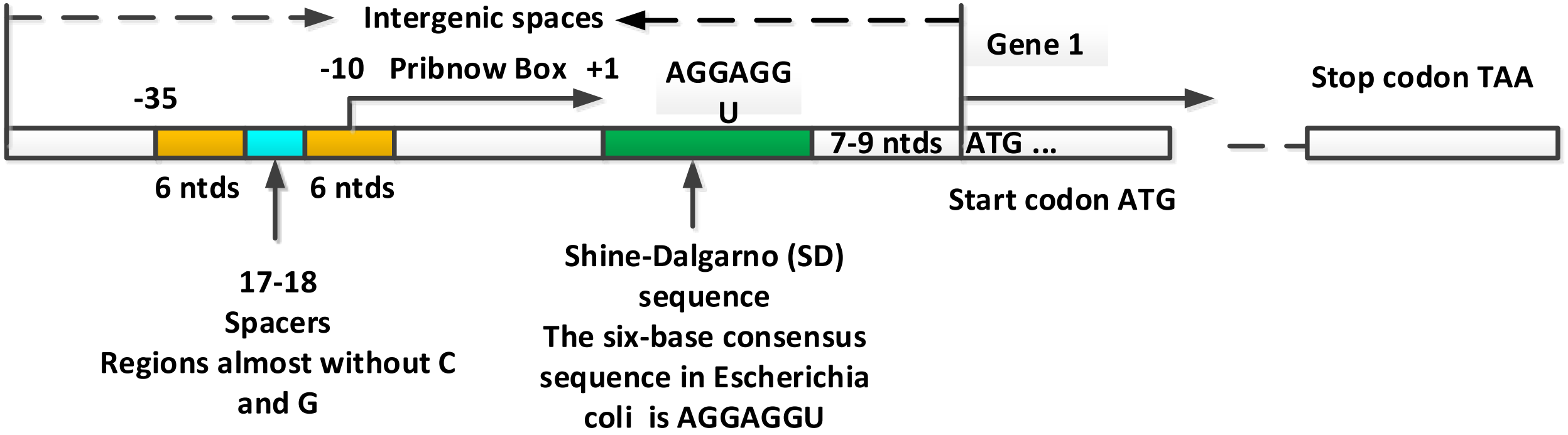
Отделните промотери често се различават от консенсусния в една или повече позиции.

PB често се нарича -10 sequence, защото е центрирана приблизително на 10 base pairs upstream от сайта на инициране на транскрипцията.

SHINE-DALGARNO (SD) SEQUENCE

- The *Shine-Dalgarno (SD) sequence* представлява свързващ сайт на рибозомата при бактериалната messenger RNA, в общия случай локализирана около 8 nts upstream от стартовия кодон AUG.
- Участва в инициирането на синтеза на протеини подравняващи рибозомата със стартовия кодон.
- Консенсусната секвенция съдържа 6 бази - **AGGAGG**
- При Escherichia coli, например, консенсусната секвенция е AGGAGGU.

СТРУКТУРА НА ПРОКАРИОТЕН ГЕН



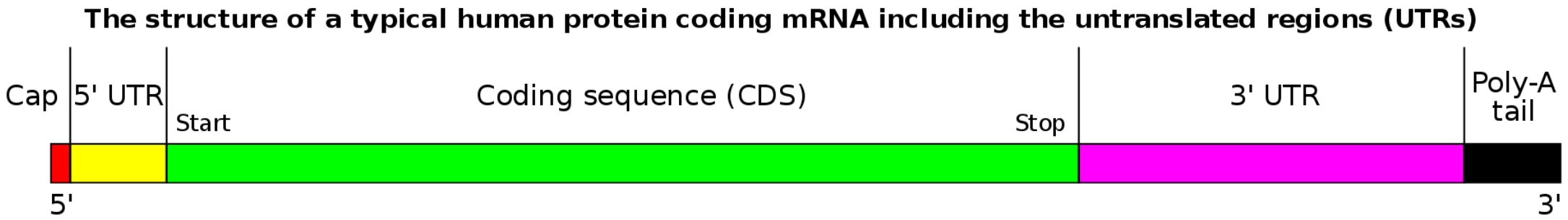
АЛГОРИТЪМ ЗА КАРТИРАНЕ НА ГЕНИ ПРИ ПРОКАРИОТИТЕ

1. Обработката започва от първия символ на секвенцията на РНК
2. От позиция -35 търсим консенсусна комбинация от 6 nts TTGACA
3. Търсим PBox от средата на втората област от 6 ntds (позиция -10).
4. От средата на PBox се отброяват 10 nts
5. От текущата позиция търсим SD областта
6. От края на SD областта се отброяват 7-9 nts и търсим стартовия кодон ATG

МЕТОДИ AB INITIO ПРИ ЕУКАРИОТИТЕ

- 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.
- 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



- Това 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

МЕТОДИ AB INITIO ПРИ ЕУКАРИОТИТЕ

- 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.

GENE FINDERS

- 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.

GENE FINDERS

- 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)

МОДЕЛ НА МАРКОВ

- Андрей Марков – руски математик (1856–1922г.), Санкт Петербургския държавен университет, Русия – известен с вериги и процеси на Марков (стохастични процеси в теория на вероятностите и статистиката)
- 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.

МОДЕЛИ 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.
- Примерно използване на верига Марков е Монте Карло верига Марков, която използва свойството Марков, за да докаже, че конкретен метод за извършване на произволна разходка ще вземе извадка от съвместното разпределение

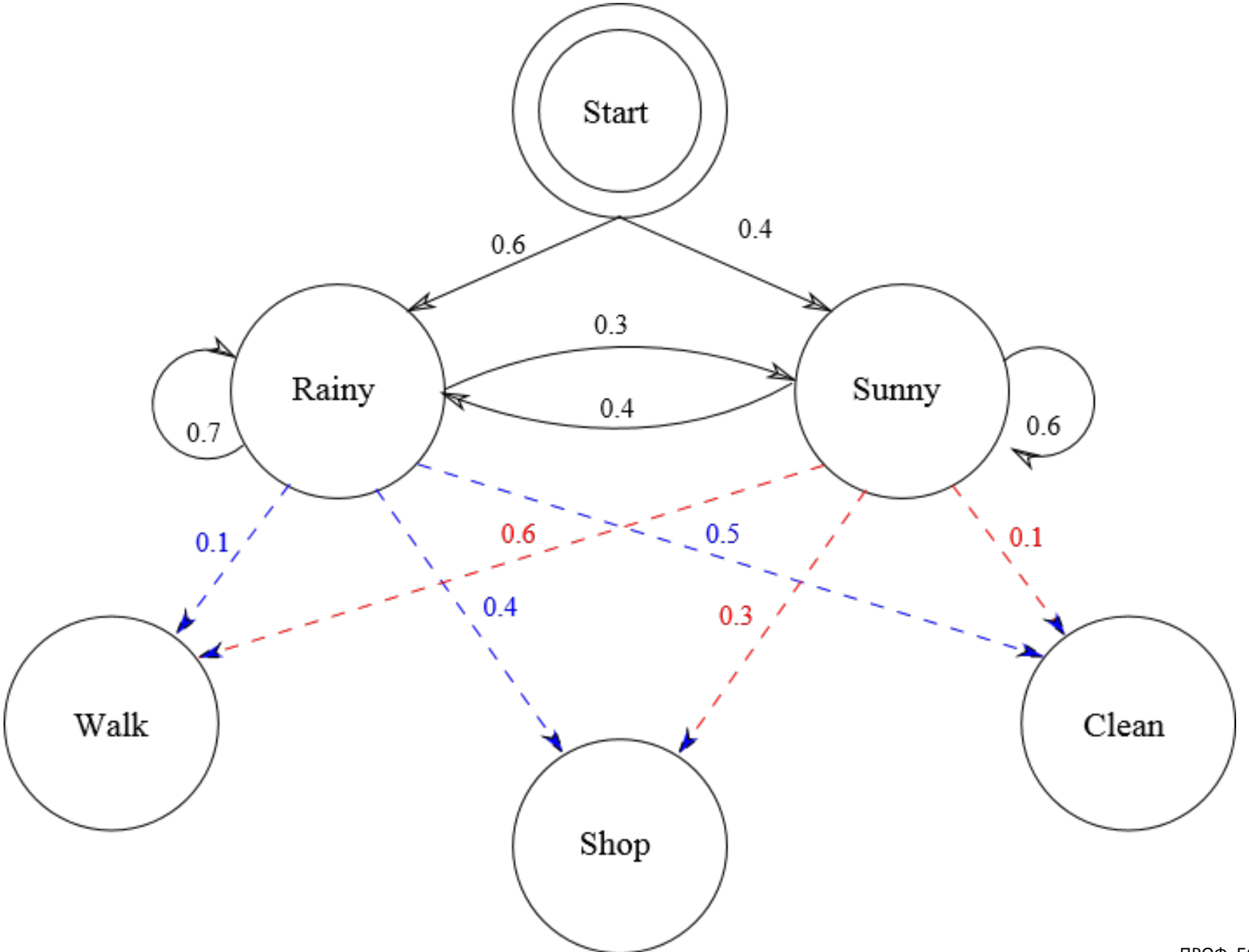
МОДЕЛИ НА МАРКОВ

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

СКРИТИ МОДЕЛИ НА МАРКОВ
HIDDEN MARKOV MODEL

- Статистически модел на Марков, при който се приема, че моделираната система е процес X на Марков с невидими (скрити – hidden) състояния
- HMM приема, че съществува друг процес Y, чието поведение зависи от процеса X
- Целта е да се получи информация за процеса X посредством наблюдение на процеса Y


Start
0.6
0.4
Rainy
Sunny
0.3
0.4
0.7
0.6
0.1
0.6
0.5
0.1
0.4
0.3
Walk
Shop
Clean

```
states = ('Rainy', 'Sunny')

observations = ('walk', 'shop', 'clean')

start_probability = {'Rainy': 0.6, 'Sunny': 0.4}

transition_probability = {
   'Rainy' : {'Rainy': 0.7, 'Sunny': 0.3},
   'Sunny' : {'Rainy': 0.4, 'Sunny': 0.6},
   }

emission_probability = {
   'Rainy' : {'walk': 0.1, 'shop': 0.4, 'clean': 0.5},
   'Sunny' : {'walk': 0.6, 'shop': 0.3, 'clean': 0.1},
   }
```

ПРИЛОЖЕНИЕ НА HMM В КОМПЮТЪРНАТА БИОЛОГИЯ

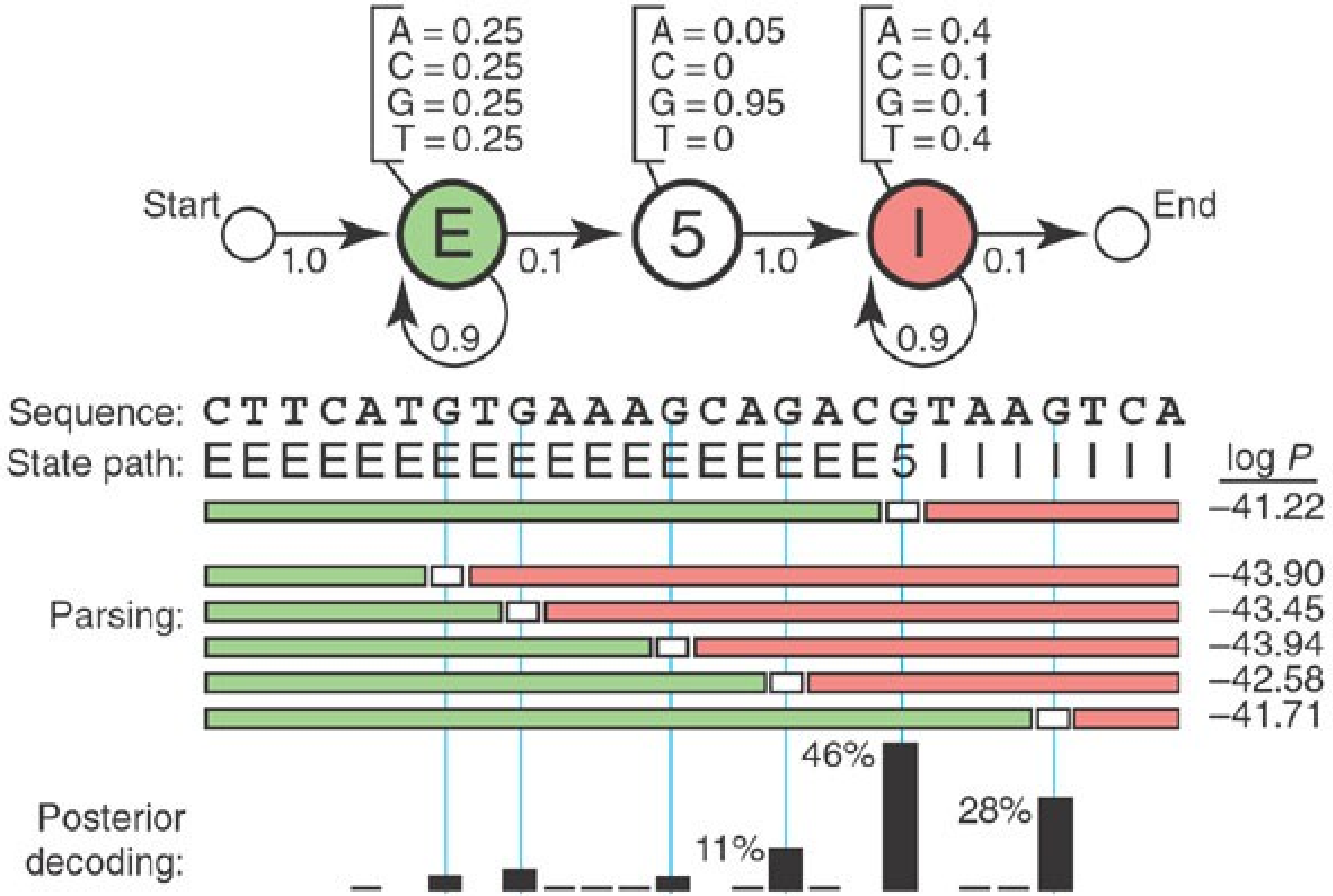
- Често анализът на биологичната секвенция е просто въпрос на поставяне на правилния етикет върху всяка аминокиселина.
- *При идентифицирането на гени целта е да се маркират нуклеотидите като екзони, интрони или интергенен участък*.
- При подравняване на секвенции целта е да се асоциират аминокиселините в секвенцията - заявка с хомоложни секвенции от целеви бази данни.
- При идентифицирането на гени (genefinder) 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 ексон/интрон, и анализ на отворената рамка за четене

ПРИЛОЖЕНИЕ НА HMM В КОМПЮТЪРНАТА БИОЛОГИЯ

- 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 са Legos на изчислителния анализ на секвенции.

ПРИМЕР: HMM ЗА РАЗПОЗНАВАНЕ НА 5′ SPLICE SITE

- Да приемем, че ни е дадена ДНК секвенция, която започва с екзон, съдържа едно 5’ сайт за слепване (splice site) и завършва в интрон.
- Проблемът е да се идентифицира къде се е стигнало до превключването от екзон към интрон - къде е 5‘ сайтът за слепване
- Приемаме, че последователностите от екзони, сплайс сайтове и интрони трябва да имат различни статистически свойства както следва:
- Екзоните имат равномерно разпределение на базите (25% за всяка база),
- Интроните изобилстват с A/T (40% за всеки A/T, 10% за всеки C/G),
- 5′ консенсус нуклеотидът е почти винаги G (95% G и 5% A).


A = 0.25
C = 0.25
G = 0.25
T = 0.25
A = 0.05
C = 0
G = 0.95
T = 0
A = 0.4
C = 0.1
G = 0.1
T = 0.4
Start
E
5
I
End
1.0
0.1
1.0
0.1
0.9
0.9
Sequence: C T T C A T G T G A A A G C A G A C G T A A G T C A
State path: E E E E E E E E E E E E E E E E E E E E E 5 I I I I I I I I
log P
Parsing:
−41.22
−43.90
−43.45
−43.94
−42.58
−41.71
Posterior decoding:
46%
28%
11%

WHAT'S HIDDEN?

- HMM генерира последователност (секвенция).
- За всяко състояние (state), определяме нуклеотид от разпределението на вероятностите в състоянието.
- След това се определя следващото състояние в съответствие с вероятностното разпределение на преходите (transitions) от текущото състояние.
- По този начин моделът генерира изходни данни под формата на два стринга
- Единият стринг представя пътя на основното състояние (етикетите - labels), тъй като се осъществява преход (transition) от едно състояние в друго.
- Другият стринг представлява наблюдаваната ДНК секвенция, като всеки нуклеотид се определя от едно състояние в пътя (последователността) на преходите.

MARKOV CHAIN

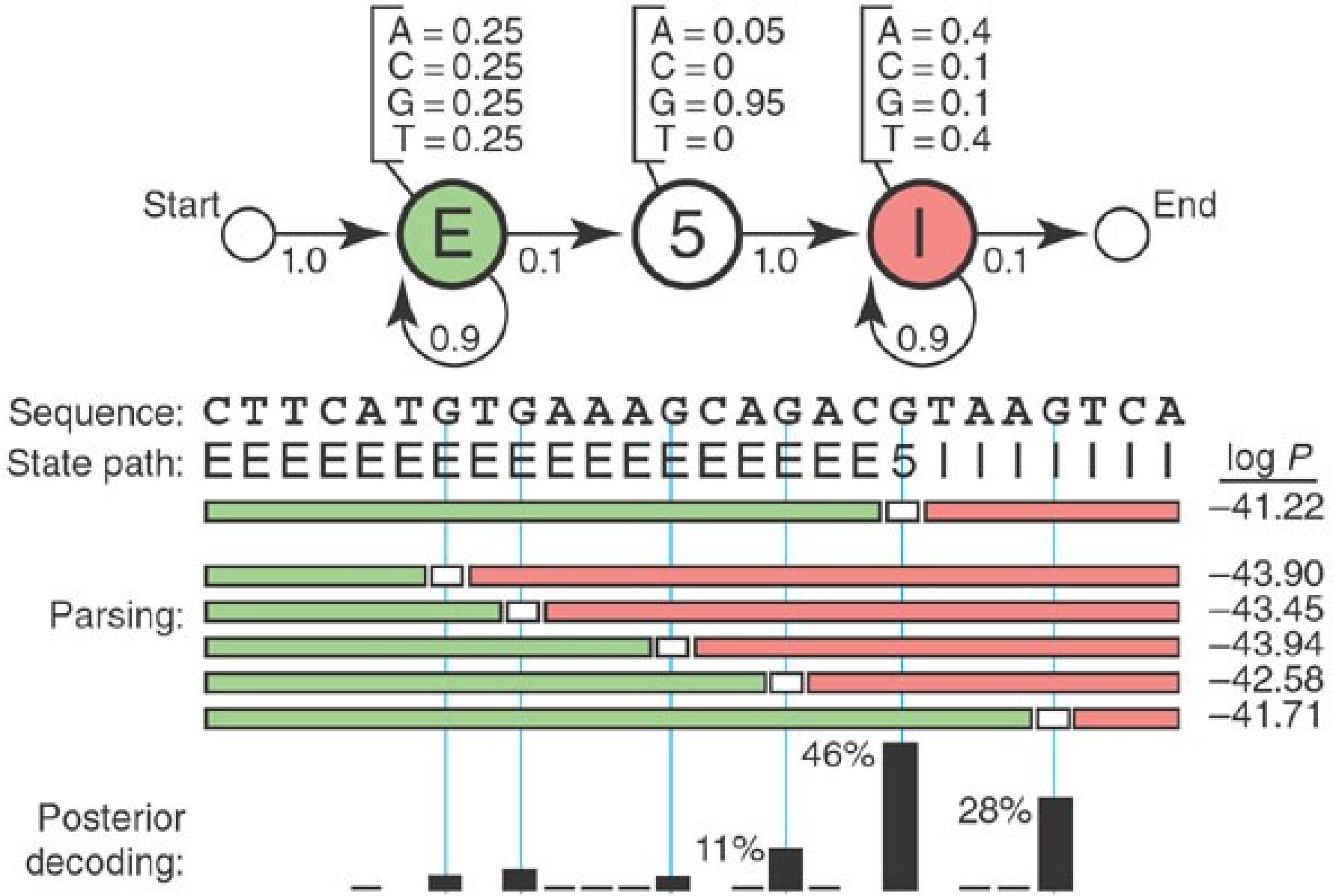
- Пътят на преходите представлява верига на Марков (Markov chain), което означава, че следващото състояние при прехода зависи само от текущото състояние.
- Поради факта, че разполагаме само с наблюдаваната секвенция, последователността на преходите е скрита – практически това са етикетите на нуклеотидите, които искаме да открием
- Пътят на преходите на състоянието представлява скрита верига на Марков (a *hidden Markov chain)*.
- *Вероятността P* (S, π | HMM, θ), че HMM с параметри θ генерира път на състоянието π и наблюдаваната секвенция S *е произведение* на всички използвани емисионни вероятности и вероятности за преход.
- В примера се наблюдава 26-нуклеотидната секвенция и пътя на състоянията в средата на фигурата, където има 27 прехода и 26 емисии, които се използват.
- Умножават се всичките 53 вероятности заедно и изчисляваме log, тъй като това са малки числа – в резултат се изчислява log P (S, π | HMM, θ) = −41.22.

HMM

- 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
- Теорема 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.

НАМИРАНЕ НА НАЙ-ДОБРИЯ ПЪТ НА ПРЕХОДИТЕ НА СЪСТОЯНИЯТА

- 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.
- Например, при зададени HMM и 26-нуклеотидната секвенция, има 14 възможни пътя, които имат ненулева вероятност, тъй като 5‘ консенсус нуклеотидът трябва да попадне върху един от 14-те вътрешни нуклеотиди A (общо 9 нуклеотида A) или G (общо 5 нуклеотида G), при вероятност 95% G и вероятност 5% A.
- В примера са изброени шестте най-добри точки за оценка (тези с G при 5 'SS).
- Най-добрият има вероятност от лог -41,22, което заключава, че най-вероятната позиция 5’SS е на петия G (46%)


A = 0.25
C = 0.25
G = 0.25
T = 0.25
A = 0.05
C = 0
G = 0.95
T = 0
A = 0.4
C = 0.1
G = 0.1
T = 0.4
Start
E
5
I
End
1.0
0.1
1.0
0.1
0.9
0.9
Sequence: C T T C A T G T G A A A G C A G A C G T A A G T C A
State path: E E E E E E E E E E E E E E E E E E E E E 5 I I I I I I I I
log P
Parsing:
−41.22
−43.90
−43.45
−43.94
−42.58
−41.71
Posterior decoding:
46%
28%
11%

АЛГОРИТЪМ НА ВИТЕРБИ (VITERBI ALGORITHM)

- 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.
- Приложения – разпознаване и синтез на говор, диаризация на говорещите лица (разделяне на входен аудио поток в хомогенни сегменти според идентичността на говорещите), идентификация на ключови думи в изказвания (keyword spotting), компютърната лингвистика, и биоинформатиката.

АЛГОРИТЪМ НА ВИТЕРБИ (VITERBI ALGORITHM)

- Създаден от Andrew Viterbi (1967г.) като алгоритъм за декодиране на конволюционни кодове при зашумени дигитални комуникационни връзки
- Неговото създаване е свързано с множество открития като алгоритъма на Needleman - Wunsch, и др.
- "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“).
- Друго приложение – проследяване на цел (target tracking), при което се определя трасето с максимална вероятност на последователност от наблюдения.

ПОСТ-ДЕКОДИРАНЕ (POSTERIOR DECODING)

- Анализ на резултата за примера – съществува алтернативен път на състоянията с логаритмична вероятност −41.71 , който се различава малко от най-добрия намерен път с логаритмична вероятност −41.22.

- Как да се уверим, че изборът на петия нуклеотид G е правилният избор?

- Предимство на вероятностното моделиране – може да определим директно сигурността (доверителността – confidence) на получените резултати.

- Вероятността нуклеотидът *i* да е емитиран (излъчен) от състояние *k* е сумата от вероятностите на всички пътища на състоянията, които обхващат състояние *k* да генерира нуклеотида *i* ($\pi_i = k$ в пътя на състоянията π), нормализирано към сумата от всички възможни пътища на състоянията.

- В примерния модел съществува само един път на състоянията в числителя, и общата сума от пътищата на състоянията е 14 в знаменателя

- За най – доброто получено решение се получава вероятност 46% за емитиране на петия нуклеотид G и вероятност 28% за емитиране на шестия нуклеотид - *posterior decoding*.

- 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.

СПЕЦИФИКА НА HMM ЗА КАРТИРАНЕ НА ГЕНИ

- 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 но партньорът в сдвоената база трябва да бъде комплементарен
- Пътят на състоянията при HMM няма опции за запомняне 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 (genefinding), 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).

SPLICE SITE CONSENSUS AND SPLICE SITE PREDICTORS

- 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** (http://ccb.jhu.edu/software/genesplicer/)

- Johns Hopkins University, Center for Computational Biology, Baltimore, Maryland
- A fast, flexible system for detecting splice sites in the genomic DNA of various eukaryotes
- **Genie: Gene Finder Based on Generalized Hidden Markov Models**
  https://www.fruitfly.org/seq_tools/genie.html