**Multiple sequence alignment of 10 genetic sequences of SARS-2 (gene 3) – T-Coffee**

**Съдържание**

1. SARS-CoV-2 - (1) (2) (3) (4)
2. Актуалност на проблемa (1) (2)
3. Описание и цел на заданието
4. Т-coffee (1) (2) (3) (4)
5. Експериментална работна рамка (1) (2) (3)
6. Експерименталните резултати (1) (2) (3)
7. **SARS-Cov-2 (1) - слайд 3 - Ники**

<https://www.mdpi.com/1420-3049/25/18/4086/htm>

<https://www.ncbi.nlm.nih.gov/sars-cov-2/>

На 11 март 2020 г. Световната здравна организация (СЗО) официално обяви огнището, причинено от новия коронавирус (SARS-CoV-2), за пандемия. Бързото разпространение на болестта изненада научната и медицинска общност и претовари здравните системи в световен мащаб. Практическите действия срещу скорошната поява и бързото разпространение на SARS-CoV-2 изискват разработването и използването на инструментите на биоинформатиката и стратегии за молекулярно моделиране, с цел откриване на потенциални анти-SARS-CoV-2 агенти.

Биоинформатиката е технология, която помага на изследователите да се справят със заболявания, като изследва генетичното секвениране и търси структурни модели на потенциални молекулярни target-и, налични в SARS-CoV-2.

Безпрецедентно огнище на пневмония с неизвестна етиология се появи през декември 2019 г. в град Ухан (провинция Хубей, Китай). Един месец по-късно учените изолират нов коронавирус (SARS-CoV-2), за който се съобщава, че е член на β(бета) подсемейството на коронавирусите. Световната здравна организация (СЗО) въведе името COVID-19 за болестта, причинена от този етиологичен агент. Предишни проучвания показват, че голям брой тежки коронавируси, свързани с остър респираторен синдром (SARSr-CoV), съществуват при прилепите и има серологични доказателства, че някои SARSr-CoV прилепи имат потенциал да заразят хората. Скорошен генетичен анализ на геномни секвенции в пълна дължина, получени от някои пациенти в ранния етап на огнището, разкриват, че секвенциите са почти идентични, споделяйки 79,6% идентичност с SARS-CoV. Освен това, SARS-CoV-2 е 96% идентичен до коронавируса при прилепите. Това посочва прилепите като вероятен произход и предполага високата възможност за предаване от животно на човек. Впоследствие беше потвърдено и предаването от човек на човек.

1. **SARS-Cov-2 (2) - слайд 4 - Ники**

По този начин броят на случаите на COVID-19 в Ухан, епицентърът на болестта, бързо се покачи и освен Китай, сега има повече от 200 държави с потвърдени случаи на COVID-19.

Фигура 1 показва броя на заразените пациенти и смъртните случаи, причинени от COVID-19 до края на август 2020 г. По отношение на огнището на SARS-CoV-2 на 30 януари 2020 г., СЗО обяви болестта за извънредна ситуация в областта на общественото здраве. Към днешна дата типичните клинични симптоми на тези пациенти са висока температура, суха кашлица, възпалено гърло, затруднено дишане (диспнея), главоболие и пневмония. Предвид експоненциалния растеж на случаите на инфекция с SARS-CoV-2, разработването на инструменти, които допринасят за откриването на нови терапевтични цели срещу вируса, е от решаващо значение за предотвратяване на повече смъртни случаи, свързани с SARS-CoV-2.

Понастоящем, при липса на ефективна терапия, известна за лечението на COVID-19 и, също така, предвид, че процесът на разработка на нови лекарства отнема много време и е тромав, биоинформатиката и молекулярното моделиране могат да бъдат използвани като инструменти, за да подпомогнат бързото развитие на лечение, което да бъде тествано в клинична среда.

Ролята на биоинформатиката във връзка с молекулярното моделиране в търсене на методи за диагностика, лечение и профилактика на COVID-19 е безспорна. Процеси като скрининг на биоактивни съединения, моделиране на биомакромолекулни структури, селекция на праймери и съединения за генетично секвениране могат да бъдат по-бързи, по-точни и по-евтини, когато се подпомагат от компютърни инструменти и автоматизация. Експерти от цял ​​свят вярват в потенциала на биоинформатиката в борбата с пандемията на COVID-19. Тя допринася за разбирането на вариациите в протеините на SARS-CoV-2 и как се увеличава вирулентността на този патоген. Въз основа на това, този инструмент(биоинформатиката) може също да изясни как вирусът подкопава имунната система.

1. **SARS-Cov-2 (3) - слайд 5 - Ники**

Повечето коронавирусни инфекции (CoV) при хората се причиняват от видове с ниска патогенност, причиняващи симптоми на обикновена настинка; обаче, в крайна сметка те могат да доведат до сериозни инфекции в по-високорискови групи като възрастни хора, пациенти със съпътстващи заболявания (хипертония, захарен диабет, астма и др.) и / или страдащи от имуносупресия. Преди 2019 г. два високопатогенни и животински коронавирусни вида (SARS и MERS) са отговорни за огнищата на тежки остри дихателни синдроми. По отношение на човешката инфекция от SARS-CoV-2, клиничният спектър не е напълно описан и смъртността, заразността и трансмисивността все още се изясняват.

Етиологичен агент

CoV, от порядъка Nidovirales, подсемейство Coronavirin и семейство Coronaviridae, е едноверижен РНК вирус с диаметър 80–120 nm, със съществуваща корона, поради наличието на гликопротеини във вирусната обвивка, която може да бъде наблюдавана под електронен микроскоп.

Това е вирус, способен да зарази хора и голямо разнообразие от други бозайници и птици, като развива заболявания на дихателната, чревната, чернодробната и централната нервна система.

CoV е класифициран в 3 подсемейства, наричани преди това групи 1, 2 и 3. Група 1 и 2 са съставени от CoV, който има бозайниците като домакини, а група 3 доскоро е била съставена само от птичи CoV. Понастоящем Изследователската група на Международния комитет за вирусна таксономия (ICVT) предложи да се заменят 3-те традиционни групи с подсемейства Alfacoronavirus (α-CoV) (група 1), Betacoronavirus (β-CoV) (група 2) и Gamacoronavirus (γ- CoV) (група 3). След това при птици и свине се открива присъствието на четвърто подсемейство CoV и се нарича Deltacoronavirus (δ-CoV).

Най-често срещаните човешки CoVs са:

* HCoV-OC43 и HCoV-HKU1 ( и двата β-COV на щам A)
* HCoV-229E и HCoV-NL63 ( и двата α-CoV)

Tе причиняват най-вече обикновени настинки и самоограничени инфекции на горните дихателни пътища при имунокомпетентни индивиди, т.е. те се елиминират за кратък период от време от имунната система, без да е необходима намесата на специфична фармацевтична терапия. Естествено при имунокомпрометирани и възрастни хора също могат да се появят инфекции на долните дихателни пътища.

Други човешки CoVs включват SARS-CoV, SARS-CoV-2 и MERS-CoV. Тези категории CoVs могат да причинят епидемии с различна клинична тежест, с респираторни и екстра-дихателни прояви. По отношение на SARS-CoV, MERS-CoV, смъртността е съответно между 10% и 35%.

1. **SARS-Cov-2 (4) - слайд 6 - Ники**

**Предаване на коронавирус между различни видове:**

(а) коронавирусът, чийто естествен преносител са прилепите (BtCoV).

Този вирус се е разпространил и се е приспособил към дивите животни, например маскирана палмова цибета, която се предлага на пазарите за морски дарове на едро в Китай, за консумация от човека. Служителите на тези пазари, които манипулират тези диви животни, са заразени. Те обаче не са показали важни клинични признаци и симптомите са минимални. Процесът на адаптиране на вируса към нови гостоприемници доведе до щамове с ефективна способност за репликация при човешки гостоприемници, които причиняват заболявания с клинични състояния, вариращи от леки до тежки и с голяма способност да се разпространяват от човек на човек.

(б) CoV-OC43 коронавирусът, чийто естествен преносител са хората и говеждият (BCoV) са тясно свързани. Предполага се, че тези коронавируси произхождат от друг животински вид и впоследствие са кръстосвали своите видове. BCoV ефективно се е разпространил сред други животински видове, например алпака (южноамерикански бозайник от семейство камили) и диви преживни животни (като елени).

1. **Актуалност на проблема (1) - слайд 8 - Деси**

Използването на биоинформатиката и други изчислителни инструменти в допълнение към молекулярното моделиране помага на изследователите от различни области при търсенето на стратегии за диагностициране на вирусната инфекция, при разработването на ваксини за нейната профилактика, както и при откриването на нови анти- CoV-2 средства.

Познаването на генома на даден вид въз основа на техниката на генетичното секвениране е отправна точка за разбирането на структурата и функцията на неговите гени. В контекста на болести, предавани от микроорганизми, като SARS-CoV-2, картографирането на генома на микроорганизми, събрани от заразени пациенти в различни региони на света, също позволява проследяване на профила на предаване, включително, неговото разпространение в различни региони и държави, допринасяне за търсене на стратегии за борба с болестта и наблюдение на мутации.

Данните от базата данни NCBI GenBank® https://www.ncbi.nlm.nih.gov/genbank), към днешна дата, показват повече от 290 000 нуклеотидни секвенции, въведени от декември 2019 г. за SARS-CoV-2, като повечето от тях идват от градове в Китай и САЩ.

През февруари 2020 г. учени от университета в Сао Пауло и Института Адолф Лутц в Бразилия, заедно с учени от университета в Оксфорд, Обединеното кралство, успяха да изяснят и публикуват пълната генна последователност на вируса, само в рамките на два дни, което е възможна благодарение на прилагането на софтуери и приложения на биоинформатиката.

1. **Актуалност на проблема (2) - 9 - Деси**

Биоинформатиката показа голям напредък през 90-те години, главно с оглед развитието на геномната област, която генерира голямо количество биологични данни, несъвместими за бърз анализ на ръка. Бързата и адекватна манипулация на тези данни е възможна чрез прилагането на софтуер за сравнение и анализ на данни и лесен достъп до наличните по-рано данни от споделянето и съхраняването на информация във виртуални бази данни. С оглед на експериментално получените последователности е възможно да се извърши подравняването на тези секвенции с други секвенции, налични във виртуални бази данни, което води до познаване на семейството, с което е свързан микроорганизмът.

1. **Описание и цел на заданието - слайд 10 - Деси**

Да се направи множествено подравняване на 10 генетични секвенции на SARS-CoV-2 (gene 3), посредством T-Coffee.

1. **T-Coffee (1) - слайд 11 - Деси**

<http://www.tcoffee.org/Projects/tcoffee/index.html#INSTALLATION>

<http://www.tcoffee.org/>

<http://www.tcoffee.org/Projects/tcoffee/papers/tcoffee.pdf>

<http://tcoffee.crg.cat/apps/tcoffee/do:regular>

Какво е T-Coffee?

T-Coffee е:

* колекция от инструменти за изчисляване, оценка и манипулиране на множествени подравнявания на ДНК, РНК, протеинови секвенции и структури. Можете да използвате T-Coffee за подравняване на секвенции или за комбиниране на резултати от други методи за подравняване (Clustal, Mafft, Probcons, Muscle ...) - в едно уникално подравняване (M-Coffee)
* нов метод за множествено подравняване на секвенции, разработен 2000г., който осигурява драматично подобрение на точността с умерена скорост в сравнение с най-често използваните алтернативи.

T-Coffee е безплатен пакет с отворен код, разпространяван под публичния лиценз на GNU.

T-Coffee може да подреди протеинови, ДНК и РНК последователности. Той също така е в състояние да комбинира информация за последователността заедно със структурна информация за протеини (3D-Coffee / Expresso), информация за профила (PSI-Coffee) или вторични структури на РНК (R-Coffee).

1. **T-Coffee (2) - слайд 12 - Ники**

<http://www.tcoffee.org/Projects/tcoffee/papers/tcoffee.pdf>

<http://www.tcoffee.org/Projects/tcoffee/index.html#INSTALLATION>

<http://www.tcoffee.org/>

<https://academic.oup.com/nar/article/39/suppl_2/W13/2505784>

T-Coffee - Tree-based Consistency Objective Function for alignment Evaluation

(целенасочена функция за оценка на съгласуваността, базирана на дърво)

Описание на алгоритъма T-Coffee:

* използва методa за прогресивно подравняване
* генерира множествено подравняване с най-високо ниво на консистентност като използва библиотека от предварително обработени глобални и локални подреждания по двойки
* може да комбинира предишни подравнения,
* може да оценява нивото на консистентност на подрежданията
* може да извлича серия от мотиви за създаването на локално подравняване

T-Coffee има две основни характеристики.

1. Първо, осигурява прост и гъвкав начин за генериране на множествени подравнявания, използвайки различни източници на данни. Данните от тези източници се предоставят на T-Coffee чрез библиотеки, които се сравняват по двойки.
2. Втората основна характеристика е оптимизационният метод, чрез който намира множеството подравнявания, които най-добре пасват на двойките подравнявания от библиотеката.
3. **T-Coffee (3) - слайд 13 - Ники**
4. Генериране на основна библиотека с подравнявания.

- тя съдържа сет от двойки измежду всички секвенции, които ще бъдат подравнявани.

- за всяка двойка секвенции се използват по два източника за подравняване - един локален и един глобален.

- Глобалните подравнявания се конструират посредством ClustalW

- Локалните са 10те топ подравнявания между всяка двойка секвенции, които се извличат посредством Lalign програмата от FASTA пакетът.

- В библиотеката всяко подравняване се представя като списък от съчетани двойки (например х от секвенция А се съпоставя с у от секвенция В). Всяка двойка има също така различна значимост, като това се определя спрямо оценъчна система.

1. T-Coffee поставя оценка (или тежест) на всяка двойка в дадената библиотека. Секвенциите се считат за акуратни, когато съвпадаемостта е повече от 30%. За всеки сет секвенции се изчисляват две главни библиотеки заедно с техните тежести/оценки. Едната посредством ClustalW глобално подравняване, а втората - локално с Lalign.
2. Целта всъщност е оптималното комбиниране на локалната и глобална информация. Това се постига чрез обединяване на двете библиотеки. Ако някоя двойка се дублира, те биват обединени в едно, а тежестта им се сумира. В противен случай се създава нова двойка. Двойки, които не са се проявили имат тежест със стойност 0 по подразбиране. Изследвайки консистенцията на всички двойки, обаче може много да увеличи стойността им, за това се назначава тежест, която рефлектира степента им на консистентност. Този

процес се нарича Library Extension.

1. Цялостната идея на екстензията е да се комбинира информацията по такъв начин, че крайната тежест за всяка двойка да рефлектира информацията от цялата библиотека.
2. Този процес от своя страна позволява използването на стратегията за Прогресивно Подравняване. Групите по двойки образуват матрица на разстоянията между секвенциите, което на свой ред образува филогенетично дърво. Най-близките двойки в това дърво се подравняват първи, посредством динамично програмиране.
3. **Експериментална работна рамка (1) - слайд 15 - Деси**

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7293463/>

Bioinfrotmatics book - p. 309

Геномът на SARS-CoV-2 се състои от едноверижна РНК и е представен в базата данни на NCBI под името NC\_045512.2 .

Вирусният геном кодира 16 неструктурни протеина, от които nsp3 е най-големият. Това е мембранен протеин с множество домейни и е от съществено значение в репликационно-транскрипционния комплекс, образуван върху мембраните на гостоприемника.

Генетичният състав на SARS-CoV-2 се състои от:

* 13–15 (12 функционални) отворени рамки за четене (ORF), съдържащи ~ 30 000 нуклеотида
* S, E, M и N протеини

Целият геном на SARS-CoV-2 кодира около 7096 остатъка с дълъг полипротеин, който се състои от много структурни и неструктурни протеини (NSP). Нуклеотидите във вирусния геном се съдържат главно в двата неструктурни протеина ORF1a и ORF1ab, последвани от структурни протеини.

1. **Експериментална работна рамка (2) - слайд 16 - Ники**

- SARS-CoV-2 -

Multi-domain non-structural protein 3 или Nsp3 - е най-големият протеин, кодиран от коронавирус геномът, с молекулна маса приблизително равна на 200 kD (килодалтона). Nsp3 представлява съществена част от репликационния комплекс.

Съставен е от различни области, 8 от които съществуват във всички познати коронавируси (CoVs). Те са следните:

* ubiquitin-like domain 1 (Ubl1)
* Glu-rich acidic domain (also called “hypervariable region”)
* macrodomain (also named “X domain”)
* ubiquitin-like domain 2 (Ubl2)
* papain-like protease 2 (PL2pro)
* Nsp3 ectodomain (3Ecto, also called “zinc-finger domain”)
* Y1 and CoV-Y of unknown functions

Триизмерните структури в N-терминала, 2/3 от Nsp3 са изследвани чрез рентгенови лъчи и ядрено-магнитен резонанс, както виждаме на изображението, а останалата част е не изследвана област, показана в червено.

Като цяло Nsp3 играе различни роли в коронавирус инфекцията. Освобождава себе си, както и Nsp1 и Nsp2 от полипротеините и взаимодейства с други Nsps, както и с РНК, за оформянето на репликационния комплекс. Напада имунната система на приемника, като същевременно се модифицира в неговите клетки посредством 3Ecto областта и подпомага оцеляването на самия вирус.

1. **Експериментална работна рамка (3), (4), (5) - слайд 17, 18, 19  
   Деси**

Етап 1, Етап 2, Етап 3

1. **Експериментални резултати (1) - слайд 21 - Деси**

**dnd** - дендрограма, генерирани от T-Сoffee във формат Newick.

- дендрограма - диаграма, представляваща дърво;

- формат Newick - в математиката, форматът на Newick дърво, е начин за представяне на теоретично графови дървета с дължини на ръбове, използвайки скоби и запетаи;

**msf\_aln, clustalw\_aln, fasta\_aln** - текстови файлове, съдържащи подравняването в различни формати. Тези файлове могат да се използва като вход за други програми.

**score\_html, score\_ascii** - цветно подравняване, появяваяващо се на фон, който показва качеството на това подравняване. Червеното - показва висококачествени сегменти. Синьото - показва региони от подравняването, на които нямате причина да вярвате. Score\_ascii е текстовата версия на .html файла. Тези два последни файла са предназначени само за показване; не можете да ги използвате като вход за други програми за анализ на секвенциите.

**score\_pdf** - pdf на цветното подравняване.

1. **Експериментални резултати (2) - слайд 22 - Деси**

**Book - p. 313**

[**https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=simple\_phylogeny-I20210505-125435-0722-16310798-p1m**](https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=simple_phylogeny-I20210505-125435-0722-16310798-p1m)

Сега, как можете да разберете дали даден блок е добър? Когато погледнете подравняването, тези три символа (\*), (:) или (.), имат много точни значения:

(\*) Звезда показва изцяло запазена колона.

(:) Двоеточие показва колони, където всички остатъци имат приблизително еднакъв размер и една и съща хидрпатия (Мярка за полярността на аминокиселинен остатък; свободната енергия за пренос на остатъка от среда с ниска диелектрична константа към вода.

(.) Точката показва колони, където има размера или хидропатията са запазени в хода на еволюцията.

Например, вашият среден добър блок е единица с дължина поне 10–30 аминокиселини, показваща поне една до три звезди (\*), още няколко двоеточия (:) близо до звездите и няколко периода (.) Поръсени тук-там.

Вълшебното при многократното подравняване на секвенции е, че 4 или 5 запазени позиции над 50 аминокиселини могат да бъдат достатъчни, за да ни убедят, че гледаме на истински сигнал. Това е по-малко от 10 процента идентичност!

Филогенетичното дърво (филограмата), показва еволюционните връзки между различни биологични видове или други образувания въз основа на сходства и разлики във физическите или генетичните им характеристики.

Предшественик, общи родове/семейство

1. **Експериментални резултати (3), (4) - слайд 23, 24 - Деси**