

Авторска справка за приносния характер на трудовете на Мая Желязкова, представени за участие в конкурс за доцент по математика /Биостатистика/

Област 4. Природни науки, математика и информатика

4.5 Математика

В4. Хабилизационен труд - научни публикации в издания, които са реферирани и индексирани в световно известни бази данни с научна информация (Web of Science и Scopus).

(Изискват се 100 точки. Представени са 2 научни публикации, носещи общо 111 точки.)

В4-1. Maya Zhelyazkova, Royamana Yordanova, Iliyan Mihaylov, Stefan Kirov, Stefan Tsonev, David Danko, Christopher Mason, Dimitar Vassilev, Origin sample prediction and spatial modeling of antimicrobial resistance in metagenomic sequencing data, Frontiers in genetics, vol:12, 2021, ISSN (print): 1664-8021, doi:doi:10.3389/fgene.2021.642991, Ref, Web of Science, IF (4.772 – 2021), Web of Science Q1 (2021), SCOPUS SJR (1.41 – 2021), SCOPUS Q1 (<https://www.scimagojr.com/journalsearch.php?q=2110>), International, (75 точки).

Резюме

Непрекъснатото развитие на проекта MetaSUB(metasub.org) поставя нови важни въпроси относно произхода, варирането и антимикробната резистентност на събраните проби от микроорганизми, които изискват нови методи за анализ на данни. Проектът CAMDA (Критичен анализ за големи масиви от данни <http://camda.info/>), в контекста на проекта MetaSUB, организира ежегодни конференции с предизвикателства, при които се тестват различни биоинформатични и статистически подходи върху проби, събрани по целия свят за бактериална класификация и

прогнозиране на географския им произход. Тази работа предлага метод, който не само предвижда местоположенията на неизвестните проби, но също така оценява относителния риск от антимикробна резистентност чрез пространствено моделиране. Въвеждаме нов компонент в стандартния анализ, докато прилагаме модел на бейсова пространствена конволюция, който отчита пространствената структура на данните, както е определено от географската дължина и ширина на пробите и оценява относителния риск от съществуването за антимикробна резистентност в регионите, което е от значение за общественото здраве. Ние можем след това да използваме изчисления относителен риск като нова мярка за антимикробна резистентност. Също така сравняваме ефективността на няколко метода за машинно обучение, като Gradient Boosting Machine, Random Forest и невронна мрежа за прогнозиране на географския произход на неизвестните проби. Трите метода показват сравними резултати с известно превъзходство на класификатора Random Forest.

Мащабните метагеномни изследвания са част от глобална инициатива за изследване и разбиране на разнообразието на микробиома. Високопродуктивни технологии за скрининг, като shotgun секвениране на цели геноми, идентифицират генетична информация на по-подробни нива, като нивото на видовете, и могат да открият и изобилието на еукариоти, гъби и вируси. Повечето методи за анализ на метагеномни последователности се базират на техники за машинно обучение с надзор. Моделите random forest често се използват за прогнозиране на географските местоположения на пробите. Повечето от тези модели са ограничени до прогнозиране на проби от местоположения, които са част от обучителните набори. За прогнозиране на нови произходи те използват Lasso регуляризация и Индекса на разнообразието на Simpson и включват предишни резултати от асоциацията между човешката генетика и географските местоположения. Наскоро бяха разработени по-сложни модели за класификация на данни за протеинови последователности, като deep learning, рекурентни и конволюционни невронни мрежи. Авторите използват различни показатели, като чувствителност, специфичност, точност, AUC, коефициент на корелация на Matthews, за да сравнят представянето на методите.

Класификацията на проби според техния произход обикновено се извършва чрез методи за наблюдавано машинно обучение, които включват разделяне на пробите на обучаващи и тестови набори. В това изследване е направен предварителен преглед на някои от добре известните методи и е решено да се използват върху три от тях, които не изискват

много параметри и са лесни за изпълнение в рамките на R. По-специално използваме Gradient Boosting Machine (GBM), Random Forest и Neural Network (NNet), както са реализирани в пакета R 3.6.3 caret. Приложени са горепосочените модели за машинно обучение, за да се предскаже на кой континент и в кой град принадлежат пробите. Обучаващите си данни са разделени на два поднабора: 60% и 40% и след това се сравняват резултатите от прогнозата върху тестовия набор. Използван е метод за избор на характеристики – Recursive Feature Elimination (RFE), който е често използван метод за селекция на характеристики, който изгражда модел и премахва най-слабите характеристики, след което тези характеристики се използват в прогнозата за континенти и градове. Тъй като антимикробната база данни има ограничен брой таксони, за частта с прогнозата се предпочита proGenomes.

В допълнение към предсказването на произхода, са приложени пространствени модели за оценка на риска от антимикробна резистентност в градовете и в различни страни. Стандартните регресивни модели, които не вземат под внимание пространствените зависимости, не работят добре тук, тъй като оценките на параметрите и резултатите ще бъдат ненадеждни. Освен това, данните показват прекомерно вариране и обикновените линейни регресионни модели ще дадат изкривени оценки. Поради това, прилагането на пространствени модели, по-специално на байсови йерархични модели, е от значение. Повече пространствена информация и вземане на проби от близко разположени градове и страни ще помогне за създаване на по-добри и по-детайлни карти на относителния риск от AMR.

Abstract

The steady elaboration of the Metagenomic and Metadesign of Subways and Urban Biomes (MetaSUB) international consortium project raises important new questions about the origin, variation, and antimicrobial resistance of the collected samples. CAMDA (Critical Assessment of Massive Data Analysis, <http://camda.info/>) forum organizes annual challenges where different bioinformatics and statistical approaches are tested on samples collected around the world for bacterial classification and prediction of geographical origin. This work proposes a method which not only predicts the locations of unknown samples, but also estimates the relative risk of antimicrobial resistance through spatial modeling. We introduce a new component in the standard analysis

as we apply a Bayesian spatial convolution model which accounts for spatial structure of the data as defined by the longitude and latitude of the samples and assess the relative risk of antimicrobial resistance taxa across regions which is relevant to public health. We can then use the estimated relative risk as a new measure for antimicrobial resistance. We also compare the performance of several machine learning methods, such as Gradient Boosting Machine, Random Forest, and Neural Network to predict the geographical origin of the mystery samples. All three methods show consistent results with some superiority of Random Forest classifier.

Large scale metagenomics studies are part of a global initiative to study and understand microbiome diversity. High-throughput screening such as shotgun whole genome sequences identifies genetic information to more detailed levels such as the level of species and can further detect abundance of eukaryotes, fungi, and viruses. Most methods for analysis of metagenomics sequence data are based on the supervised machine learning techniques. Random forest models are often used in predicting geographical locations of the samples. Most of those models are limited to predicting samples from locations that are part of the training sets. For predicting new origins used Lasso regularization and Simpson's diversity index and incorporated previous results of association between human genetics and geographical locations. Recently more complex models have been developed for classification of protein sequence data such as deep learning, recurrent and convolution neural networks. Authors used different measures such as sensitivity, specificity, accuracy, AUC, Matthews correlation coefficient to compare the performance of the methods. The classification of samples by their origin is commonly performed by supervised machine learning methods which involve dividing the samples into training and testing sets. In this work we did preliminary review of some of the well-known methods and decide to focus on three of them that do not involve many parameters and are easy to run within the framework of R. In particular we use Gradient Boosting Machine (GBM), Random forest, and Neural network (NNet) as implemented in the R 3.6.3 package caret. We applied the above machine learning models to predict which continent and which city the samples belong to. We split our training data into two subsets: 60% and 40% and then compare the prediction results on the test set. To avoid the batch effect we analyze the samples separately by the sampling day. Recursive feature elimination (RFE), a commonly used feature selection method that fits a model and removes the weakest features, is used to screen for the top features which are then used in the prediction of continents and cities. Since

the antimicrobial database has a limited number of taxa, proGenomes is preferred for the prediction part of analysis.

In addition to the prediction of origins we apply spatial models to access the risk of antimicrobial resistance inside cities and across countries. Standard regression models that do not take into account the spatial dependencies do not work well here since the parameter estimates and results will be unreliable. Moreover, the data show overdispersion and ordinary linear regression models will produce biased estimates. Therefore, applying spatial models in particular Bayesian hierarchical models is relevant. More spatial information and sampling of closely located cities and countries will help to build better and more detailed maps of AMR relative risk.

B4-2. Maya Zhelyazkova, Roumyana Yordanova, Iliyan Mihaylov, Stefan Tsonev, Dimitar Vassilev, *In silico* discovering relationship between bacteriophages and antimicrobial resistance, *Biotechnology and Biotechnological Equipment*, vol:37, issue:1, 2023, pages:14-23, ISSN (online): 1310-2818, doi:10.1080/13102818.2022.2151378, Ref, IF, IF (1.762 - 2021), Web of Science Quartile: Q4 (2023) (36 точки).

Резюме

Бактериофагите и техният потенциален принос към антимикробната резистентност привличат нарастващо внимание в контекста на медицината и фармацевтиката. Основната цел на предизвикателството CAMDA е да се придобият повече знания за връзката между вирусите, техните гостоприемници и гените за антимикробна резистентност при определяне на факта дали антимикробната резистентност наистина може да се разпространява чрез фаги. Това изследване е фокусирано върху връзката и възможните зависимости между бактериофагите и антимикробна резистентност на основа на данните, събрани от повечето градски среди по света. Пробите в данните са класифицирани в две категории: висока и ниска, според наблюдаваните нива на гени за антимикробна резистентност (AMR). Подходът, използван в нашия анализ, се състои от няколко различни метода, които оценяват различното изобилие на фаги, тяхното разнообразие в пробите, въздействието им върху категории антимикробна резистентност и асоциации с антимикробни резистентни гени. Връзката между фагите, техните гостоприемници и антимикроб-

ната резистентност също е изследвана чрез бейсов пространствен модел, използвайки категорията за антимикробна резистентност (нисък срещу висок) като фактор. Установяваме по-висок относителен риск за фаги, които инфектират *Staphylococcus aureus* самостоятелно или едновременно със *Staphylococcus aureus* и *Acinetobacter baumannii* в групата с висока AMR, което предполага, че тези фаги може да имат роля в разпространението на антимикробни гени. Новото в тази разработка е приложението на Бейсовия пространствен модел.

Идеята за метагеномни изследвания се основава на поколението от огромни набори от данни от кратки последователности на четене от прокариотни клетки в микробни общности. Антимикробните профили показват регионални различия в глобална градска канализация и Subway и Urban среди. Някои проучвания също изследват връзка между разнообразието на бактериофагите и ARG профили. Много от съществуващите инструменти за анализ метагеномните последователности разчитат на така наречения k-mer подход за извличане на микробни таксони от такива данни. Някои скорошни методи са разработени въз основа на подходи за задълбочено обучение за намиране на потенциални мутации или вирусни последователности.

Предизвикателството CAMDA2021 предоставя набор от 124 проби от среди на метрото и градския обществен транспорт по всички континенти, разделени на две групи въз основа на техните ARG профили (62 висок и 62 нисък AMR проби). Брой на оперативната таксономична единица (OTU), таксони на фаги, нормализирани като прочитания на милион (RPM) бяха извлечени от последователностите на shotgun sequences на проби, използващи Kaiju - v1.8.2 метагеномен класификатор с вирусната база данни на Националния център за Биотехнологична информация (NCBI). Класификаторът използва класифициране на ниво протеин и постига висока чувствителност. Общият брой на ненулевите изобилни вируси е 6618 докато общият брой на ненулевите изобилни фаги е 2687. Също така бяха извлечени бактериални таксони, използвайки proGenomes база данни и беше обобщен разчита на нивото на вида. За AMR изобилието бяха използвани данните от профила на ARG, предоставени от CAMDA и Rangea, които включват нормализирани множества както на ARG, така и на класове таксони.

Една от новостите на тази разработка е приложението на CoDa анализ. Композиционните данни (CoDa) са вектори на неотрицателни компоненти, показващи относителното тегло или важност на набор от ча-

сти в общата сума. Частите не могат да се тълкуват изолирани една от друга. Стандартни методи като корелационен анализ, който предполага Евклидовата геометрия, приложена директно към композиционни данни, може да доведе до пристрастия в резултатите, тъй като примерното пространство за композиционни данни е симплекс. Един от често срещаните подходи за анализиране на композиционни данни е да се прилагат стандартни методи за логаритмично съотношение трансформирани данни и след това използвайте обратна трансформация на логаритмично съотношение, за да се върнете към оригиналното пространство. Данните се считат за композиционни, ако се състоят от вектори с положителни компоненти, чиято сума е постоянна (например 1 за пропорции или 100 за проценти). По-конкретно, това дефинира примерното пространство като хиперравнина, наречена Aitchison Simplex

$$S^D = \left\{ x = (x_1, x_2, \dots, x_D) \in R_+^D \mid \forall k > 0 \exists! \lambda > 0 : x = \lambda C_k(x) \right\}.$$

Метагеномните данни (напр. данни за броя на OTU) са композиционни данни, тъй като сборът от всички стойностите на компонентите зависят от процедурата за вземане на проби и данните са пропорционални. Като вземем предвид тази структура, ние извършваме диференциален анализ на изобилието използвайки композиционни методи. Сравнени са изобилието на фагите за високо и ниско AMR проби, за да се установи за кои фаги има разлика между двете групи. Използван е R-пакета ALDEx2, който е разработен с помощта на Бейсови методи за откриване на ANOVA тип диференциални разлики специално за композиционни данни. Също така е извършен точен тест на Fisher, за да се открие възможно свръхпредставяне на семействата на таксони на фаги.

Ще дефинираме Алфа разнообразие с индекси Chao1, Shannon, Simpson и Beta разнообразие с индекс на несходство на Брей-Къртис. Алфа разнообразието измерва вариациите на микробите в една проба докато бета разнообразието измерва вариациите между пробите. В тази работа се използват и двете мерки, за да се характеризират данни. По-специално се използва три индекса като Chao1 индекс на разнообразие, който е определен като

$$S_{chao1} = S_{obs} + \frac{n_1^2}{2n_2},$$

където S_{chao1} е изчисленият брой видове, S_{obs} е общият брой видове наблюдавани в извадката, n_1 е броят на видовете, представени само от един индивид и n_2 е броят на видовете, представени от два индивида.

След това се използва Shannon индекс, който е дефиниран по следния начин:

$$H = - \sum_{i=1}^S p_i \ln p_i.$$

В тази формула p_i е делът на общността, представена от i -тия вид, а S е общият брой на видовете. И двата индекса Chao1 и Shannon придават по-голямо значение към по-рядко срещаните видове. Накрая се използва индекса на Simpson, който представя вида хетерогенност и се определя като

$$D = 1 - \sum_{i=1}^n p_i^2,$$

където p_i е относителното изобилие на i -тия вид в пробата. Simpson индексът варира от 0 (ниско разнообразие) до почти 1. Противно на описаните индекси по-горе, индексът на Simpson дава по-голяма тежест на по-доминиращите видове. Индексът на несходство на Bray-Curtis е статистическа мярка, използвана за количествено определяне на композиционно различие между две различни проби, базирано на преброяване във всяка проба и се определя по следния начин:

$$BC = \frac{\sum_{i=1}^n |X_{ij} - X_{ik}|}{\sum_{i=1}^n (X_{ij} + X_{ik})},$$

където X_{ij} , X_{ik} е броят на индивидите във вида i във всяка проба (j, k) и n е общият брой видове в пробите. Използват се горните методи за оценка на видовото богатство, а именно Chao 1, индексите на разнообразие на Shannon-Wiener и Simpson като внедрени в пакета R веган. Също се извършва тест на Kruskal-Wallis за сравняване на мерките за разнообразие през AMR статус и географско местоположение. Прилагаме се ограничен анализ на принципа с несходството на Брей-Къртис (capscale от R пакет веган). Също се използва PERMANOVA (функция адонис във веган пакета), която разделя сумите на квадратите, използвайки полуметрични и метрични матрици на разстоянието. Той извършва подобни на ANOVA тестове на вариацията в бета разнообразието, което се обяснява с категорни предиктори като състояние на AMR или географско местоположение. Методът обикновено се прилага при големи генетични данни като метагеномни OTU данни.

Моделът LASSO се прилага към данни за фаги за предсказване на вариацията на AMR в рамките на пробите и впоследствие за отчитане на

двата основни класа на AMR – висок AMR и нисък AMR. Нека обозначим AMR класове по G , като се вземат стойностите 1 и 2. Ще се опише модел на логистична регресия, която представя класовите условни вероятности чрез линейна функция на предикторите

$$Pr(G = 1|x) = \frac{1}{1 + \exp(-(\beta_0 + x'\beta))}.$$

Алтернативно, това предполага

$$\log \frac{Pr(G = 1|x)}{Pr(G = 2|x)} = \beta_0 + x'\beta.$$

Разделят се данните на набор за обучение и набор за тестване и се използва кръстосано валидиране, както е предвидено в Lasso пакета `saret`.

Анализът включва и Бейсово пространствено моделиране. Пробите в предизвикателството CAMDA2021 имат пространствена информация като географска ширина и дължина и предишни проучвания показват, че метагеномните профили на таксоните показват пространствени корелации, затова използваме бейсов йерархичен модел, по-специално Besag–York–Mollie, който е конволюционен модел с CAR априорно разпределение. Моделът използва матрица W за пространствено съседство, която измерва разстоянието между проби въз основа на наличната географска информация като географска ширина и дължина. Моделът ще оцени рискът на фага, който е последваща оценка на SIR съотношение (наблюдавано/очаквано) на фаг в проби, където очакваното изобилие е пропорционално към броя на съответните таксони на бактерии гостоприемници. Моделът включва AMR категория (ниска срещу висока) като потенциален фактор. Ако относителният риск на фага е по-висока от 1 и надеждният набор от категория AMR го прави да не включва нула, можем да заключим, че изобилието на фага е по-високо от това, което бихме очаквали въз основа на броя на бактериите гостоприемници и този по-висок риск може частично да се обясни с антимикробната категория състояние. Например една AMR група (напр. висока) има повишен риск, докато другата група (напр. нисък) има намален риск спрямо очакваните числа които са пропорционални на бактерията гостоприемник на фага. Съответните бактерии гостоприемници за всеки фаг бяха измерени от <https://www.genome.jp/virushostdb/>. В случай, когато фагите заразяват повече от една бактерия се включва допълнителното изобилие от бактерии гостоприемници в модела. (За описание на модела в по-строги математически термини, вижте Lawson.)

$$O_i|E_i, \theta_i \sim \text{Poisson}(E_i\theta_i), \quad i = 1, \dots, n$$

$$\ln(\theta_i) = \beta_0 + \nu_i + u_i$$

Тук ν са пространствено неструктурирани случайни ефекти които приемат нормално разпределение, докато u са произволни ефекти, които улавят пространствената автокорелация между пробите. Тогава

$$u_i|u_j, j \neq i, \quad W \sim N\left(\frac{1}{n_i} \sum_{i \sim j} u_j, \frac{\sigma_u^2}{n_i}\right).$$

Използваме $W = 1 - D$ (нормализирано разстояние), което е съседната матрица между позициите за географска ширина и дължина на пробите. Приема се, че откликът следва разпределението на Поасон и той отчита свръхдисперсията $Var(O) > E(O)$ и това е предимство пред чистия модел на Поасон. Ние използваме изпълнението на бейсовите настройки в пакета R 3.6.3 CARBayes, където изводът се основава на Марков верига Монте Карло симулация (MCMC). За измерване на пространствената автокорелация е използван I тестът на Moran. Сходимостта на модела също се проверява от Geweke z-scores. Ние допълнително изчисляване на корелациите между фагите и специфични ARG.

Анализът на CoDa идентифицира фаги, показващи различно изобилие между групи с висока и ниска AMR. По-голямата част от тези фаги принадлежат на Семейство Myoviridae.

Използвайки машинно обучение, идентифицирахме фаги, които могат да се използват като предиктори на AMR статус: идентифицираните таксони принадлежат главно към Myoviridae и семейство Siphoviridae. Пространственият бейсов модел показва по-висок относителен риск от изобилие на фаги в групата с висока AMR, когато се контролира изобилието на гостоприемника, което показва че фагите могат да играят роля в разпространението на ARG.

Abstract

Bacteriophages and their potential contribution to antimicrobial resistance have attracted growing attention in the context of medicine and pharmaceuticals.

A major objective of the CAMDA challenge is to acquire more knowledge about the relationship between viruses, their hosts and antimicrobial resistance genes in determining if antimicrobial resistance indeed can spread through phages. This study is focused on exploring the relationship and possible dependencies between bacteriophages and antimicrobial resistance based on the data collected from urban environments all over the world. The samples in the data are classified into two categories: high and low, according to the observed levels of antimicrobial resistance (AMR) genes. The approach used in our analyses consists of several different methods which assess the differential abundance of phages, their diversity across samples, the impact on antimicrobial resistance categories and associations with antimicrobial resistance genes. The relationship between phages, their hosts and antimicrobial resistance is also explored by a Bayesian spatial model, using the antimicrobial resistance category (low vs high) as a factor. We found a higher relative risk for phages known to infect *Staphylococcus aureus* alone or both *Staphylococcus aureus* and *Acinetobacter baumannii* in the high AMR group, which implies that these phages may have a role in the dissemination of antimicrobial genes. The new approach in this methodology is the application of Bayesian spatial model.

The idea of metagenomic studies is based on generation of huge datasets of short read sequences from prokaryotic cells in microbial communities. Antimicrobial profiles exhibit regional differences in global urban sewage and Subway and Urban environments. Some studies also explored the relationship between bacteriophage diversity and ARG profiles. Many of the existing tools for analyzing metagenomic sequences rely on the so-called k-mer approaches to extract microbial taxa from such data. Some recent methods have been developed based on deep learning approaches to find potential mutated or viral sequences.

The CAMDA challenge provides a set of 124 samples from Subway and Urban public transport environments across all continents divided into two groups based on their ARGs profiles (62 high and 62 low AMR samples). Operational Taxonomic Unit (OTU) counts of phages taxa normalized as reads per million (RPM) were extracted from the shotgun sequences of the samples using Kaiju-v1.8.2 metagenomics classifier with the virus database of the National Center for Biotechnology Information (NCBI). The classifier uses protein level classification and achieves high sensitivity. The total number

of non-zero abundant viruses is 6618 while the total number of non-zero abundant phages is 2687. We also extracted bacteria taxa using proGenomes2 database and summarized the counts on the level of species. For the AMR abundances we used the ARG profile data provided by CAMDA and Pangea which included normalized sets of both ARGs and taxa classes.

One of the new methods applied in this paper is the CoDa analysis. Compositional data (CoDa) are vectors of non-negative components showing the relative weight or importance of a set of parts in a total. Parts cannot be interpreted isolated from the others. Standard methods such as correlation analysis, which assumes Euclidean geometry applied directly to compositional data, may lead to biased results since the sample space for compositional data is simplex. One of the common approaches of analyzing compositional data is to apply standard methods to log-ratio transformed data and then use inverse log-ratio transformation to return to the original space. Data are considered compositional if they consist of vectors with positive components whose sum is constant (e.g. 1 for proportions or 100 for percentages). More specifically, this defines the sample space as a hyperplane called Aitchison Simplex

$$S^D = \left\{ x = (x_1, x_2, \dots, x_D) \in R_+^D \mid \forall k > 0 \exists! \lambda > 0 : x = \lambda C_k(x) \right\}.$$

Metagenomics data (e.g. OTU count data) are compositional data since the sum of all component values depends on the sampling procedure and the data are proportional. Taking into account this structure, we perform differential abundance analysis using compositional methods. We compare the phages abundance for high and low AMR samples to find for which phages there is a difference between the two groups. We use the R-package ALDEx2, which was developed using Bayesian methods for detecting ANOVA type of differential expressions specifically for compositional data. We also perform Fisher exact test to find possible overrepresentation of families of phages taxa.

Finally, we define Alpha diversity with Chao1, Shannon, Simpson indices and Beta diversity with Bray-Curtis dissimilarity Index. Alpha diversity measures variation of microbes in a single sample whereas beta diversity measures variation across samples. For more detailed information about the applications of diversity measures in metagenomic analysis and the methods for their calculation. In this work we use both measures to characterize our data. In particular we use three indices such as Chao1 diversity index, which

is defined as

$$S_{chao1} = S_{obs} + \frac{n_1^2}{2n_2},$$

where S_{chao1} is the estimated number of species, S_{obs} is the total number of species observed in the sample, n_1 is the number of species represented by only one individual and n_2 is the number of species represented by two individuals.

Next, we use the diversity index, which is defined as

$$H = - \sum_{i=1}^S p_i \ln p_i,$$

where p_i is the proportion of the community represented by the i th species and S is the total number of species. Both Chao1 and Shannon indices give more importance to less common species. Finally, we use Simpson index, which presents the species heterogeneity and is defined as

$$D = 1 - \sum_{i=1}^n p_i^2,$$

where p_i is the relative abundance of the i th species in the sample. Here Simpson's index ranges from 0 (low diversity) to almost 1. Contrary to the indices described above, Simpson index gives more weight to the more dominant species. The Bray-Curtis dissimilarity index is a statistic measure used to quantify the compositional dissimilarity between two different samples, based on counts at each sample and is defined as follows

$$BC = \frac{\sum_{i=1}^n |X_{ij} - X_{ik}|}{\sum_{i=1}^n (X_{ij} + X_{ik})},$$

where X_{ij} , X_{ik} is the number of individuals in species i in each sample (j, k) and n is the total number of species in samples. We use the above methods for estimation of species richness, namely Chao 1, Shannon-Wiener and Simpson diversity indices as implemented in the R package *vegan*. We perform Kruskal-Wallis test to compare the diversity measures across AMR status and geographical location. We apply Constrained Analysis of Principal Coordinates with Bray-Curtis dissimilarity (capscale from R package *vegan*). We also use PERMANOVA (function *adonis* in the *vegan* package), which partitions sums of squares using semi-metric and metric distance matrices.

It performs ANOVA-like tests of the variance in beta diversity which is explained by categorical predictors like AMR status or geographical location. The method is usually applied to large genetic data such as metagenomics OTU data.

In this study, we apply several types of regression models provided in R package *caret* and we focused on Lasso (Least Absolute Shrinkage and Selection) penalized regression for high-dimensional data. This model adds a L1 type penalty term (which is estimated from data by cross-validation technique) for some variables in the model which are highly correlated. As a result, the penalty term will be able to set up some of the model coefficients exactly as zero. Thus, the LASSO is a penalized type of the model selection approach. The estimated model is more parsimonious and interpretable than the models obtained with other regression models like the ridge regression. The LASSO model has many applications because it is an efficient method in regression problems that arise in many scientific areas. The LASSO model is applied to the phages data to predict the AMR variation within the samples and subsequently to account for the two major classes of AMR – High AMR and Low AMR. Let us denote the AMR classes by G taking the values 1 and 2. As described before, the logistic regression model represents the class-conditional probabilities through a linear function of the predictors

$$Pr(G = 1|x) = \frac{1}{1 + \exp(-(\beta_0 + x'\beta))}.$$

Alternatively, this implies that

$$\log \frac{Pr(G = 1|x)}{Pr(G = 2|x)} = \beta_0 + x'\beta.$$

We split the data into a training and a testing set and use cross-validation as provided in the Lasso implementation of package *caret*.

Next, we apply Bayesian spatial modeling. Since the samples in CAMDA-2021 challenge have spatial information such as latitude and longitude and previous studies indicated that metagenomics taxa profiles exhibit spatial correlations, we use a Bayesian hierarchical model, in particular Besag–York–Mollie, which is a convolution model with a CAR prior. The model uses a spatial neighbourhood matrix W which measures the distance between the samples based on the available geographic information such as latitude and longitude. The model will estimate the phage’s risk which is a posterior

estimate of the SIR ratio (Observed/Expected) of a phage across the samples where the expected abundances are proportional to the corresponding host bacteria taxa counts. The model includes the AMR category (low vs high) as a potential factor. If the phage’s relative risk is higher than 1 and the AMR category credible set does not include zero and we can conclude that the phage’s abundance is higher than what we would expect based on the host bacteria counts and this higher risk can be partially explained by the antimicrobial category status. For example, one AMR group (e.g. high) has an increased risk while the other group (e.g. low) has a decreased risk relative to the expected numbers which are proportional to the phage’s host bacteria. The corresponding host bacteria for each phage were queried from

<https://www.genome.jp/virushostdb/>

In case where phages infect more than one bacteria we include the additional host bacteria abundance in the model. (For a description of the model in more strict mathematical terms, see Lawson.)

$$O_i|E_i, \theta_i \sim \text{Poisson}(E_i\theta_i), \quad i = 1, \dots, n$$

$$\ln(\theta_i = \beta_0 + \nu_i + u_i$$

Here ν are spatially unstructured random effects that assume normal distribution while u are the random effects that capture the spatial autocorrelation between samples. Then

$$u_i|u_j, j \neq i, \quad W \sim N \left(\frac{1}{n_i} \sum_{i \sim j} u_j, \frac{\sigma_u^2}{n_i} \right).$$

We use $W = 1 - D$ (normalized Distance) which is the neighboring matrix between samples latitude and longitude positions. The response is assumed to follow Poisson distribution and it accounts for overdispersion $Var(O) > E(O)$ and this is an advantage over the pure Poisson model. We use the Bayesian setting implementation in R 3.6.3 package CARBayes, where inference is based on Markov chain Monte Carlo simulation (MCMC). The model is fit with the function S.CARbym from the above package. Moran’s I test was used to measure the spatial autocorrelation. The model convergence is also checked by Geweke z-scores . We further calculate the correlations between phages and specific ARGs.

The CoDa analysis identified phages showing different abundances between high and low AMR groups. The majority of these phages belonged to the Myoviridae family. Using machine learning, we identified phages that could be used as predictors of AMR status: the identified taxa belonged mainly to the Myoviridae and Siphoviridae families. The spatial Bayesian model showed a higher relative risk of phage abundances in the high AMR group when controlling for host abundance, which indicates that phages may play a role in ARG dissemination.

Област 4. Природни науки, математика и информатика

4.5 Математика

Г7. Научни публикации, които са реферирани и индексирани в световно известни бази данни (Web of Science и Scopus), извън хабиляционния труд.

(Изискват се 200 точки, представени са 7 научни публикации, носещи общо 255 точки.)

Г7-1. Ivanova-Aleksandrova, Nadya, Trifonova, Iva, Panayotova, Elitsa, Dimitrov, Dimitar, Marinov, Martin, Gladnishka, Teodora, Taseva, Evgenia, Ivanova, Vladislava, Zhelyazkova, Maya, Kamenov, Galin, Zehtinjev, Pavel, Christova, Iva, *Borrelia burgdorferi* Infection in Bird-Feeding and Questing Ticks in Bulgaria, *Acta Zoologica Bulgarica*, Volume 76, Issue 3, 425-430, September, 2024, ISSN 03240770, <https://www.scopus.com/record/display.uri?eid=2-s2.0-85206994675>, SCOPUS Quartile Q4, (36 точки).

Резюме

Изследвано беше наличието на *Borrelia burgdorferi sensu lato* (причинител на Лаймска болест) в кърлежи *Ixodes ricinus* отстранени от диви птици и при гладни кърлежи. Общо 345 кърлежа от 75 опаразитени птици принадлежащи към 10 вида и 274 гладни кърлежи, събрани от околната растителност, бяха изследвани чрез PCR за *Borrelia burgdorferi* s. l. ДНК на патогена е намерена в 8,1% (28/345) от хранещите се с птици кърлежи

– 10,9% (11/101) в ларвите и 7,0% (17/244) в нимфите. Разпространението на инфекцията е 9,4% при кърлежи събрани от кос (*Turdus merula*), 4,0% при кърлежите от пойния дрозд (*T. philomelos*) и 16,6% от червеногръдката (*Erithacus rubecula*). Откриване на заразени ларви (13,3%), събрани от кос предполага, че този вид може да допринесе за предаването на инфекцията към вектора *I. ricinus*. Общото разпространение на *B. burgdorferi* в нимфи е 10,6%. Това е първият доклад за заразяване с *B. burgdorferi* s. l. на кърлежи *I. ricinus* от диви птици в България.

За да се оцени епидемиологичната роля на коса като резервоар сравнихме процент на заразяване с *B. burgdorferi* s. l. между нимфите от косове и гладните нимфи с помощта на интерактивен калкулатор (Preacher 2001) за χ^2 (хиквадрат) тест. Изчисленията са направени със степени на свобода $(2-1) * (2-1) = 1$ и р-стойност $< 0,05$ беше избрана като критична стойност на нивото на значимост.

Abstract

The presence of *Borrelia burgdorferi sensu lato* (the agent of Lyme disease) in *Ixodes ricinus* ticks removed from wild birds and questing ticks was investigated. A total of 345 ticks from 75 infested birds belonging to 10 species and 274 questing ticks collected from the surrounding vegetation were examined by PCR for *Borrelia burgdorferi* s. l. Pathogen DNA was found in 8.1% (28/345) of the bird-feeding ticks – 10.9% (11/101) in larvae and 7.0% (17/244) in nymphs. Prevalence of infection was 9.4% in ticks from the Eurasian blackbird (*Turdus merula*), 4.0% in ticks from the song thrush (*T. philomelos*) and 16.6% from the European robin (*Erithacus rubecula*). Detection of infected larvae (13.3%) collected from the Eurasian blackbird suggests that this species might contribute to the transmission to the vector *I. ricinus*. The overall prevalence of *B. burgdorferi* in the questing nymphs was 10.6%. This is the first report of *B. burgdorferi* s. l. infection in *I. ricinus* from wild birds in Bulgaria.

To assess the epidemiological role of the blackbirds as reservoirs, we compared the *B. burgdorferi* s. l. infection rate between the nymphs from the Eurasian blackbirds and the questing nymphs using an interactive calculator (Preacher 2001) for χ^2 (chisquare) test. Calculations were made with degrees of freedom $(2-1) * (2-1) = 1$, and a p-value < 0.05 was chosen as the critical value of the significance level.

Г7-2. Hodzhev Y, Tsafarova B, Tolchkov V, Youroukova V, Ivanova S, Kostadinov D, Yanev N, Zhelyazkova M, Tsonev S, Kalfin R, Panaiotov S, Visualization of the individual blood microbiome to study the etiology of sarcoidosis, *Comput Struct Biotechnol J*. 2023 Oct 20;22:50-57. doi: 10.1016/j.csbj.2023.10.027. PMID: 37928975; PMCID: PMC10624578, IF-4.5, Q1, (75 точки).

Резюме

Единичните микробни патогени и по-рядко дисбиозата на микробиома на гостоприемника се спрягат като основни причинители на фиброзни и грануломатозни белодробни заболявания, например саркоидоза. Метагеномното секвениране предоставя възможност за анализ на микробиомния състав в неговата цялост. Получените от този тип данни са известни с това, че са сложни по структура и ненормално разпределени. Допълнителна трудност идва от факта, че повечето грануломатозни белодробни заболявания са с ниска честота и събирането на големи набори от пациентски данни представлява значително предизвикателство. Поради тази причина данните за единични пациенти често могат да бъдат обобщени само качествено, посредством различни методи за визуализация. Изследването на малки извадки надгражда базовите статистически методи и дава допълнителни възможности за статистически извод.

Целта на това изследване е да се извлече максимално надеждна информация за количествения и качествен състав на кръвния микробиом на гостоприемника за целите на диагностиката и клиниката посредством таргетно метагеномно секвениране на кръвни проби от единични пациенти и малки извадки.

Метагеномните резултати на седем пациенти с белодробна саркоидоза бяха сравнени с тези на 22 здрави лица. Общата микробна ДНК беше извлечена от лизирана кръв и подложена на таргетно 16S секвениране (секвенира се специфичен фрагмент от генът за 16S рибозомната субединица, включващ регионите V3 - V4). Почистените от артефакти данни бяха таксономично идентифицирани с помощта на софтуера Kraken 2. Индивидуалните метагеномни данни бяха визуализирани чрез диаграми на Sankey, диаграми на Krona и heat map. Бяха идентифицирани пет рода, които присъстват изключително или значително засилено при пациенти със саркоидоза: Veillonella, Prevotella, Cutibacterium, Corynebacterium и Streptococcus.

Предложеният подход позволява характеризиране на състава и разнообразието на микробиома в кръвта при редки заболявания на индивидуално ниво и на ниво малки пациентски извадки. Изследването на кръвния микробиом при пациенти с грануломатозни белодробни заболявания с неизвестна етиология, като саркоидоза, може да допринесе за по-добро разбиране на произхода и патогенезата им. Това потенциално би довело до откриването на нови персонализирани терапевтични подходи.

Уеб-базираната услуга MicrobiomeAnalyst (достъпна на 9 януари 2023 г.) беше използвана за статистически анализ на микробиомния състав и многообразие при пациенти със саркоидоза в сравнение с контролната група. Услугата използва пакета MicrobiomeAnalyst R.

За оценка на сходството на микробния състав между пробите от една извадка — алфа-разнообразие, което се характеризира с нормално разпределение. Беше изчислен индексът на многообразие на Shannon. Груповите вариации в алфа-разнообразието бяха оценени с параметричен t -тест, а нормалното разпределение на данните беше проверено чрез теста на Kolmogorov-Smirnov. Методът за анализ на главните координати (PCoA) е приложен за оценка на междугруповото бета разнообразие (пациенти спрямо контролна група), използвайки индекса на несходство на Bray-Curtis.

Нормализацията на данните се осъществи чрез разреждане (rarefaction) за справяне с вариациите в дълбочината на секвениране, когато броят на последователните сканирания в метагеномните библиотеки надхвърля 10 пъти. Йерархичното клъстериране използва евклидова мярка за разстояние и алгоритъма на Ward. Индивидуалното нормализирано изобилие на ниво род беше представено чрез heat map.

Настоящото изследване показва, че диаграмите на Sankey и Krona ефективно илюстрират таксономичното разнообразие на кръвния микробиом както на индивидуално, така и на групово ниво. Според нас диаграмите на Sankey са по-интуитивни и подходящи за визуализации. Установихме повишено присъствие на пет рода: Veillonella, Prevotella, Cutibacterium, Corynebacterium и Streptococcus. Въпреки това, тяхната роля в развитието на саркоидоза и формирането на белодробни грануломи остава не напълно изяснена, но подобно увеличение може да се счита за пилотен диагностичен маркер.

Необходими са допълнителни проучвания за потвърждаване на тези резултати и за внедряването им в клиничната практика. В бъдеще създа-

ването на бази данни с метагеномни, клинични и демографски данни ще подпомогне идентифицирането на биомаркери и прагове за по-ефективна диагностика.

Такъв анализ се прави за първи път в България.

Abstract

Single microbial pathogens and, less commonly, dysbiosis of the host microbiome are implicated as major causes of fibrotic and granulomatous lung diseases, such as sarcoidosis. Metagenomic sequencing provides the opportunity to analyze the microbiome composition in its entirety. The data obtained from this type of analysis are notoriously complex in structure and aberrantly distributed. An additional difficulty comes from the fact that most granulomatous lung diseases are of low incidence, and collecting large patient data sets is a significant challenge. For this reason, data from single patients can often only be summarized qualitatively, using various visualization methods. The study of small samples extends the scope of descriptive statistics and provides additional opportunities for statistical inference.

The aim of this study is to extract the most reliable information about the quantitative and qualitative composition of the host's blood microbiome for diagnostic and clinical purposes through targeted metagenomic sequencing of blood samples from single patients and small samples.

Metagenomic results from seven patients with pulmonary sarcoidosis were compared with those from 22 healthy individuals. Total microbial DNA was extracted from lysed blood and subjected to targeted 16S sequencing (a specific fragment of the 16S ribosomal subunit gene, encompassing the V3-V4 regions, is sequenced). Artifact-cleaned data were taxonomically identified using Kraken 2 software. Individual metagenomic data were visualized using Sankey plots, Krona plots, and heat maps. Five genera were identified as being exclusively or significantly overrepresented in patients with sarcoidosis: *Veillonella*, *Prevotella*, *Cutibacterium*, *Corynebacterium*, and *Streptococcus*.

The proposed approach allows for the characterization of the composition and diversity of the blood microbiome in rare diseases at the individual level and in small patient samples. The study of the blood microbiome in patients with granulomatous lung diseases of unknown etiology, such as sarcoidosis, may contribute to a better understanding of their origin and pathogenesis. This could potentially lead to the discovery of new personalized therapeutic approaches.

The web-based service MicrobiomeAnalyst (accessed January 9, 2023) was used to statistically analyze the microbiome composition and diversity in sarcoidosis patients compared to controls. The service uses MicrobiomeAnalyst R package.

To assess the similarity of microbial composition between samples from a single sample — alpha diversity, which is characterized by a normal distribution — the Shannon diversity index was calculated. Group variations in alpha diversity were assessed with a parametric t-test, and the normal distribution of the data was checked with the Kolmogorov-Smirnov test. Principal coordinates analysis (PCoA) was applied to assess between-group (patients vs. controls) beta diversity using the Bray-Curtis dissimilarity index.

Data normalization was performed by rarefaction to address variations in sequencing depth when the number of consecutive reads in the metagenomic libraries exceeded 10-fold. Hierarchical clustering used Euclidean distance measure and Ward's algorithm. Individual normalized abundance at the genus level was represented by a heat map.

Our study shows that Sankey and Krona diagrams effectively illustrate the taxonomic diversity of the blood microbiome at both the individual and group levels. In our opinion, Sankey diagrams are more intuitive and suitable for visualization. We found an increased presence of five genera: Veillonella, Prevotella, Cutibacterium, Corynebacterium and Streptococcus. However, their role in the development of sarcoidosis and the formation of pulmonary granulomas remains incompletely understood, but such an increase can be considered as a pilot diagnostic marker.

Further studies are needed to confirm these results and to implement them in clinical practice. In the future, the creation of databases with metagenomic, clinical and demographic data will help identify biomarkers and thresholds for more effective diagnostics.

The novelty is that this is the first time such analysis is done in Bulgaria.

Г7-3. Maya Zhelyazkova, Roumyana Yordanova, Maroussia Slavtchova-Bojkova, A Bayesian spatial analysis of mumps data in Bulgaria, *Pliska Studia Mathematica Bulgarica*, vol:27, 2017, pages: 123-130, ISSN (print):0204-9805, ISSN (online): 2367-5535, Ref, Zentralblatt (1424.62179), (18 точки).

Резюме

Бейсовите пространствени методи са широко прилагани в различни научни области като епидемиологични изследвания, обработка на изображения и много други. В тази работа е използван Бейсов йерархичен модел с Гаусово условно авторегресионно априорно разпределение към седмични данни за епидемията от заушка от 2007 г. в България. Генерирана е карта на заболяването на суровия стандартизиран коефициент на заболяемост във всички регионални центрове. Създава се и подобно картографиране за изгладен относителен риск. Комбинацията от методи за оценка на относителният риск е мощен инструмент за идентифициране на региони с по-висок риск и може да бъде използвана за насочване на местните власти и програми.

Пространственият автокорелационен анализ е полезен инструмент за анализиране на минали огнища и подпомагане на здравните власти за разбиране на пространственото разпределение на такива огнища във времето.

През 2007 г. в България имаше епидемия от заушка. От 1 януари до 18 март са регистрирани 997 случая от 2007 г. Засяга най-вече по-младите поколения между 15-19 и 20-24 години. Основната причина за епидемията от паротит е лошата имунизация при тези възрастови групи. За да се справят с епидемията, здравните власти решиха да предложат допълнителна имунизация срещу заушка. Данните за заушка в това изследване са избрани от първите 12 седмици на 2007 г., събрани от 28 различни региона на България. За всеки регион имаме 12 наблюдения, които представляват новите случаи на заушка в последователни седмици на 2007 г. Данните са на Националния център на Инфекциозни и паразитни болести. Основната ни цел е да анализираме повторно данните за заушката за да получим по-добро разбиране на епидемията.

Картографирането на заболяванията търси области с повишен относителен риск. Има 28 регионални центрове и 262 общини в България. Означаваме с $O = (O_1, O_2, \dots, O_n)$ наблюдаваният брой случаи на заушка във всеки областен център, $n = 28$ за първите 12 седмици на 2007 г. Изчисляваме очаквания брой случаи E_i , $i = 1, \dots, 28$ като умножим броя на хората във всеки областен център по заболяемостта на заушка в България. Процентът на заболяемост е броят на случаите на заушка за първите 12 седмици на 2007 г., разделени на населението на България. Ние получаваме съотношение на честота $SIR_i = O_i/E_i$ за всеки регион. Коефициентът на заболяемост е груба оценка на относителния риск. След това търсим региони с повишен риск. Ако относителният риск е

по-голям от единица, тогава регионът е изложен на по-висок риск от болестта. Ако относителният риск е по-малък от единица, популацията е сравнително здрава. Понякога рискът се повишава случайно поради малък E_i , защото заболяването е рядко или населението в района е малко. Ето защо използваме Бейсови йерархични модели, които намаляват вероятността от повишаване на риск случайно.

Конволюционния метод ни дава по-добра оценка на разпространението на заушката, като това е използвано иновативно в дадената работа. В световен мащаб е също иновативен. Може да се използват в геостатистиката и анализ на данни от геоложки проучвания.

Конволюционният модел отчита влиянието на съседните области върху резултата от избраната област, който е по-близо до реалната картина при изучаване на разпространение на инфекциозна болест. Това се вижда от анализа на заболяемостта в различни области на страната. Картографирането на SIR е грубо приближение и понякога е трудно за тълкуване. Ето защо се разглежда гладката версия, която е картографиране на относителния риск като моделна оценка от конволюционния модел. От друга страна гладкото картографиране в някои случаи може да доведе до свръхизглаждане и няма да разкрие всички характеристики на относителния риск.

Abstract

Bayesian spatial methods have been widely applied in different scientific areas such as epidemiological studies, image processing and many others. In this work we use Bayesian hierarchical model with Gaussian conditionally autoregressive prior to a collection of weekly mumps data from 2007 outbreak in Bulgaria. We generate a disease mapping of the crude standardized incidence ratio across all regional centers. Similar mapping is also produced for the smoothed relative risk. The combination of methods for estimates of the relative risk is a powerful tool to identify high risk regions and may be used to guide local authorities and programs.

Spatial autocorrelation analysis is a useful tool to analyze past outbreaks and help health authorities understand the spatial distributions of such outbreaks over time.

There was an outbreak of mumps in Bulgaria in 2007 as discussed in Kojouharova et al. There were 997 cases registered between January 1st and March 18th of 2007. It affected mostly younger generations between ages 15-19 and 20-24. The main reason for the mumps outbreak was the poor immunisation in these age groups. To deal with the outbreak, the health authorities decided to offer a supplementary mumps immunisation. The mumps data in this study are selected from the first 12 weeks of 2007 collected from 28 different regions of Bulgaria. For every region we have 12 observations which represent the new cases of mumps in consecutive weeks of 2007. The data is provided by the National Center of Infectious and Parasitic Diseases. Our main goal is to reanalyze the mumps data in order to get better understanding of the outbreak.

Disease Mapping looks for areas with elevated relative risk. There are 28 regional centres and 262 municipalities in Bulgaria. We denote by $O = (O_1, O_2, \dots, O_n)$ the observed number of cases of mumps in each regional center, $n = 28$ for the first 12 weeks of 2007. We calculate the expected number of cases $E_i, i = 1, \dots, 28$ by multiplying the number of people in each regional center by the incidence rate of mumps in Bulgaria. The incidence rate is the number of cases of mumps for the first 12 weeks of 2007 divided by the population of Bulgaria. We obtain an incidence ratio $SIR_i = O_i/E_i$ for each region. The incidence ratio is a crude estimate of the relative risk. We then search for regions with elevated risk. If the relative risk is greater than one, then the region is exposed to a higher risk of the disease. If the relative

risk is less than one, the population is comparatively healthy. Sometimes, the risk is elevated by chance due to small E_i because the disease is rare or the population in the area is small. That is why we use Bayesian hierarchical models which decrease the probability of having elevated risk by chance.

The convolution method allows us to better estimate the spread of mumps, and this has been used innovatively in the given work. On a global scale it is also innovative. It can be used in geostatistics and analysis of data from geological surveys.

The convolution model takes into account the influence of neighbouring areas on the outcome of the selected area, which is closer to the real picture in studying the spread of infectious disease. This is evident in the analysis of incidence in different areas of the country. The SIR mapping is a crude approximation and sometimes is hard to interpret. That is why we consider a smooth version which is disease mapping of the relative risk as a model estimate from the convolution model. On the other hand the smooth mapping in some cases can produce oversmoothing and it will not reveal all the features of the relative risk.

Г7-4. D. Yovchev, E.Deliverska, J.Indjova, Maya Zhelyazkova, Mandibular incisive canal: A cone beam computed tomography study, Biotechnology and Biotechnological Equipment, vol:27, issue:3, 2013, pages:3848-3851, ISSN (print):1310-2818, Ref, IF, IF (0.31 - 2013), Web of Science Quartile: Q4 (2013), SCOPUS Quartile: Q4 (2013), (36 точки)

Резюме

Целта на това проучване е да се оцени наличието, ширината и началната част на мандибуларния инцизивен канал (МІС) и неговата връзка с менталния форамен (МF), с помощта на конусовидна компютърна томография (СВСТ). Материалът на изследването включва 140 СВСТ сканирания, направени за предоперативно планиране на поставянето на импланти. Внимателно са проучени СВСТ данни от трима опитни наблюдатели за наличието и местоположението на началната част на МІС и връзката му с MF. Вътрешният диаметър на МІС в началото на канала е измерен три пъти от един наблюдател. Съгласието между наблюдателите при идентифицирането на МІС беше отлично (средна капа стойност от 0,84). МІС е открит при 92,9 % от случаите. Средният вътрешен

вертикален диаметър на канала е $1,44 \text{ mm} \pm 0,39 \text{ mm}$. Диаметърът на канала достига максимум $2,5 \text{ mm}$. Каналите при мъжете и тези от дясната страна на долната челюст са значително по-широки. По-голямата част от МИС – 86,3 % ($n = 195$), започват преди МФ (тип 1); останалите – 13,7 % ($n = 31$), започват изключително близо до форамена (тип 2). МИС е почти постоянна находка при СВСТ сканиране. В случаите, когато МИС започва на нивото на менталния форамен (тип 2) лингвалното позициониране на имплантите може да спаси невроваскуларния сноп в канала. Всичко това ще доведе до по-добри условия за остеоинтеграция и предотвратяване на евентуални невросензорни смущения.

Всички данни бяха събрани и статистически анализирани от SPSS 16.0 (SPSS Inc., Чикаго, Илинойс, САЩ). Р-стойности от 0,05 или по-малко се считат за статистически значими. Споразумение между наблюдателите беше постигнато чрез решение на мнозинството (поне двама от трима наблюдатели са съгласни). Капа статистиката беше приложена за оценка на споразумение между наблюдателите. Капа коефициентът се интерпретира като лошо (0), слабо (0,01–0,20), сравнително (0,21–0,40), умерено (0,41–0,60), значителни (0,61–0,80) и почти перфектни (0,81–1,0), според Ландис и Кох. По отношение на измерването на диаметъра на инцизивния канал, вариациите в рамките на наблюдателя бяха изразени чрез коефициентите на вариация (CV). За оценка беше проведен двустранен независим t -тест за възможни разлики между половете и сдвоен t -тест за определяне на разликите между диаметрите от двете страни на долната челюст.

Abstract

The aim of this study was to assess the presence, width, and initial part of the mandibular incisive canal (MIC) and its relation with the mental foramen (MF), by means of cone beam computed tomography (CBCT). The study material included 140 CBCT scans taken for preoperative planning of implant placement. CBCT data were examined carefully by three experienced observers for the presence and location of the initial part of the MIC and its relation with the MF. The inner diameter of the MIC at the beginning of the canal was measured three times by one observer. The interobserver agreement in identification of the MIC was excellent (mean kappa value of 0.84). A MIC was found in 92.9 % of cases. The mean inner vertical diameter of the canal was $1.44 \text{ mm} \pm 0.39 \text{ mm}$. The diameter of the canal reached 2.5 mm at

maximum. The canals in males and these on the right side of the mandibles were significantly wider. The majority of the MICs –86.3 % ($n = 195$), began before the MF (type 1); the rest of them – 13.7 % ($n = 31$), began extremely close to the foramen (type 2). The MIC is an almost permanent finding on CBCT scans. In cases when the MIC begins at the level of the mental foramen (type 2) the lingual positioning of the implants could save the neurovascular bundle in the canal. All this will lead to better conditions for osteointegration and prevent eventual neurosensory disturbances.

Data and statistical analysis All data were gathered and statistically analyzed by SPSS 16.0 (SPSS Inc., Chicago, Illinois, USA). P-values of 0.05 or less were considered statistically significant. Agreement between observers was reached by means of a majority decision (at least two of three observers agreed). Kappa statistics were applied for assessment of interobserver agreement. The kappa coefficient was interpreted as being poor (0), slight (0.01–0.20), fair (0.21–0.40), moderate (0.41– 0.60), substantial (0.61–0.80), and almost perfect (0.81–1.0), according to Landis and Koch. Regarding the measurement of the incisive canal diameter, the intraobserver variations were expressed by the coefficients of variation (CV). A two-tailed independent *t*-test was carried out to assess possible differences between genders, and a paired *t*-test to determine differences between the diameters on the two sides of the mandible.

Г4-5. Maya Zhelyazkova, Roumyana Yordanova, Iliyan Mihailov, Stefan Tsonev, Dimitar Vassilev, Bioinformatics and Biostatistical Models for Analysis and Prognosis of Antimicrobial Resistance, Recent Contribution to Bioinformatics and Biomedical Sciences and Engineering, editor/s:Lecture Notes in Networks and Systems, Publisher:Springer, 2023, pages:53-62, ISBN:378-3-031-31068-3, doi:10.1007/978-3-031-31069-0-7, Ref, SCOPUS, SJR (0.151 - 2021), (30 точки)

Резюме

Връзката между бактериофагите и антимикробната резистентност може да бъде преразгледана в обхвата на потенциалното въздействие на микробиологията върху съвременната медицина и фармацевтика. Досега изследванията даваха двусмислени резултати чрез проучване и потвърждаване на влиянието на бактериофагите върху промяната на ан-

тимикробни резистентни гени. Основната цел на нашата разработка е да се придобият нови знания как вирусите, техните гостоприемници и антимикробните резистентни гени са свързани и как тези отношения могат да бъдат изяснени в контекста на разпространение на антимикробни резистентни гени чрез фаги. Представената работа е ориентирана към изясняване на връзката и възможните зависимости между бактериофаги и антимикробна резистентност. Данните са събрани от различни среди по целия свят. Настоящият анализ се състои от различни биоинформатични и биостатистически методи за оценка на различното изобилие от фаги, тяхното разнообразие в пробите, въздействие върху категориите антимикробна резистентност и асоциациите с антимикробни резистентни гени.

Новата част от нашия анализ е използването на бейсов йерархичен модел, който отчита географското местоположение на пробите и оценява относителния риск на бактериофагите, като включва информация за техните бактериални хостове. Бактериофагите могат да предават гените за антимикробна резистентност чрез бактериална трансдукция или предаването на ARG може да се осъществи чрез бактериална конюгация без участието на бактериофаги. Във втория случай се очаква изобилието на фагите да е пропорционално на изобилието на хостовете. Бейсовият пространствен анализ ни помага да тестваме хипотезите за ролята на фагите в разпространението на ARG, като изследваме как фагите еволюират в пространствено корелирани проби, използвайки както изобилието на фаги, така и на хостове. Този модел ще бъде много полезен, когато имаме тясно свързани общини и градове, за да можем да проследим как фагите еволюират и дали относителният риск се променя през границите. Това е много подобно на модела за картографиране на заболявания, но в този контекст вирусът е бактериофаг, а човешкият хост е съответстващият бактериален хост.

Abstract

The relationship between bacteriophages and antimicrobial resistance could be regraded in the scope of potential impact of microbiology on contemporary medicine and pharmaceuticals. Till now research projects have some ambiguous outcomes by examining and confirming the influence of bacteriophages on variation of antimicrobial resistance genes. The major goal of our study is to acquire a new knowledge how the viruses, their hosts and antimicrobial

resistance genes are related and how these relations can be clarified in the context of antimicrobial resistance dissemination through phages. The presented work is oriented towards zooming in the relationship and possible dependencies between bacteriophages and antimicrobial resistance. The data has been collected from different city environments all over the world. Our analyses consist of different bioinformatics and biostatistical methods for assessment of the differential abundance of phages, their diversity across samples, the impact on antimicrobial resistance categories and associations with antimicrobial resistant genes.

The novel part of our analysis is the use of a Bayesian hierarchical model to take into account the geographical location of the samples and estimate the relative risk of the phages by incorporating information about their host bacteria. The bacteriophages can transmit the antimicrobial resistance genes through bacterial transduction or the ARG transmission can be achieved through bacterial conjugation without the involvement of bacteriophages. In the second case, the phages abundance is expected to be proportional to the host abundance. Bayesian spatial analysis help us to test the hypotheses about the role of the phages in the dissemination of ARG by examining how the phages evolved across spatially correlated samples using both phages and hosts abundances. This model will be very useful when we have closely related municipalities and cities so we can track how the phages evolve and if the relative risk changes across borders. This is very similar to the disease mapping model but in this setting the virus is a bacteriophage and the human host is the corresponding bacteria host. For improving the accuracy of the model, we can further use several additional factors as potential covariates: climate conditions and antibiotic usage.

Г4-6. Maya Zhelyazkova, Roumyana Yordanova, Iliyan Mihaylov, Stefan Kirov, Stefan Tsonev, David Danko, Dimitar Vassilev, Bayesian Hierarchical Modelling for Antimicrobial Resistance , Contemporary Methods in Bioinformatics and Biomedicine and their Applications, Lecture Notes in Networks and Systems, Publisher:Springer, 2022, pp: 79-87, ISSN (print):2367-3370, ISSN (online):2367-3389, ISBN:978-3-030-96637-9, doi:<https://doi.org/10.1007/978-3-030-96638-6>, Ref SCOPUS, SJR (0.17-2021), International, (30 точки).

Резюме

Антимикробната резистентност представлява сериозен здравен проблем. Разработени са изчислителни и статистически подходи за откриване и класификация на антимикробна резистентност. Технологиите на секвениране от следващо поколение ни позволяват да анализираме сложни метагеномни съобщества, както и наличието на резистоми в тях. В тази работа е приложен бейсов йерархичен пространствен модел за оценка на относителния риск от свързани с антимикробна резистентност таксони чрез използване на наличната пространствена информация на проби.

CAMDA2020 набор от данни е последователности от данни за пробите, взети на 21 юни през 2016 и 2017 година. За всеки град има около 50 проби, разпределени в различни зони на метрото. Всяка проба е секвенирана с средно 6 милиона двойни четения с дължина 125bp, използвайки Illumina NGS секвенсори. Участниците имат достъп до повече от 1000 проби с двойни четения от 23 града на различни континенти. Метаданни за всяка проба и за всеки град включват пространствена информация, данни за времето (температура, налягане, валежи, влажност) и демографски данни (население, плътност на населението, вид на населеното място).

Първичната обработка на данните е с помощта на Trimmomatic v0.39 с следните параметри: минимална дължина от 50 и средно качество в прозорец с размер 3, което не е по-малко от 20. Базата данни за антимикробна резистентност от NCBI съдържа последователности на гени, свързани с антимикробни резистентни микроорганизми. Всеки от гените е свързан с конкретен микробен вид (OTUs), като OTU означава оперативна таксономична единица. Съответните последователности във формат fasta се използват в класификатора Kaiju за метагеномика. Общият брой на видовете, които притежават таксони на антимикробна резистентност (гените) в изследването, е 443. Допълнително трансформираме суровите бройове в RPKM (reads/kilobase/million mapped reads). Тъй като данните са взети в различни дни (една година по-късно), при тяхното разглеждане наблюдаваме силен ефект на партидите и затова решихме да анализираме данните поотделно за всеки ден.

Пространствените модели, като конволюционния модел на Besag-York-Mollié, са полезни за оценка на риска от антибиотична-резистентност в даден град и между различни градове и страни. Картите на Фигури 1 и 2, както и сравнението между градовете в Фигури 3 и 4, могат да бъдат много полезни за определяне на области с по-висок относителен риск за специфични резистентни микроорганизми. Увеличаването на обема

на извадката чрез събиране на повече проби по целия свят ще помогне за изграждането на по-добри и по-детайлни карти на относителния риск от антибиотични-резистентни таксони. Бейсовите йерархични модели ще отчетат така нареченото "прекомерно вариране което често се наблюдава при метагеномни данни. Освен това, можем да използваме нулево-насочен поасонов модел (смесен модел от разпределение, базирано в нулата, и поасоново разпределение) за броя, когато има излишък от нули. Допълнителни пространствени модели, като тези на Leroux et al., Lee and Mitchell и Lee and Sarran, които разглеждат локално-гладки методи, също могат да бъдат приложени. В заключение, е предложен оценения относителен риск като начин за характеризиране на антибиотична-резистентност, която може да бъде полезна в приложения, свързани с общественото здраве.

Abstract

Antimicrobial resistance poses a serious health problem. Computational and statistical approaches have been developed for bacterial antimicrobial resistance discovery and classification. Next generation sequencing technologies allows us to analyze complex metagenomes and the presence of resistomes in them. In this work we apply Bayesian hierarchical spatial model to estimate the relative risk of antimicrobial resistance related taxa by using the available spatial information of the samples.

CAMDA2020 dataset is sequence data for the samples taken on June 21st in 2016 and 2017. For each city there are about 50 samples spread across different areas of the subway system. Each sample was sequenced with an average of 6M 125bp paired-end reads using Illumina NGS sequencers. Participants have access to more than 1000 samples with pair reads from 23 cities across different continents. Meta data for each sample and for each city include spatial information, weather data (temperature, pressure, precipitation, humidity) and demographics (population, population density, type of settlement). We preprocess the data using Trimmomatic v0.39 with the following parameters: minimum length of 50 and the average quality in a window size of 3 is no less than 20. The antimicrobial resistance database from NCBI consists of sequences of genes related to antimicrobial resistant microorganisms. Each of the genes is associated with a specific microbial species (OTUs), where OTU stands for operational taxonomic unit. The corresponding fasta format sequences are then used in Kaiju metagenomics

classifier. The total number of species possessing antimicrobial resistance taxa (genes) in the study was 443. We further transformed the raw counts to RPKM (reads/kilobase/million mapped reads). Since the data are taken on different days (a year apart) when we examined them we observe a strong batch effect and therefore decided to analyze the data on each day separately.

Spatial models such as the convolution Besag-York-Mollie model are helpful in assessing the risk of Antimicrobial resistance inside a city and across cities and countries. The maps in Figure 1 and 2 and the comparison among cities in Figures 3 and 4 can be very useful to determine areas of higher relative risk for specific resistant microbes. Collecting more samples across the globe will help to build better and more detailed maps of relative risk of antimicrobial resistant taxa. Bayesian hierarchical models will account for such overdispersion which is commonly observed in metagenomics data. In addition, we can use a zero-inflated Poisson model (mixture of a point mass distribution based at zero and a Poisson distribution) for counts with excessive zeros. Additional spatial models such as Leroux et al., Lee and Mitchell and Lee and Sarran which consider locally smooth methods can also be applied. In conclusion we propose the estimated relative risk as a way to characterize the Antimicrobial resistance which may be useful in applications related to public health.

G4-7. Ira Dimitrova, Ivan Dimov, Maya Zhelyazkova, E-learning course "Physical and Colloidal Chemistry" at the universities, Proceedings of the 11-th International Conference on Virtual Learning, ICVL, October 29-th, 2016, Publisher: Bucurest University Press, 2016, pages 99-103, ISSN(print):1844-8933, Ref WoS (30 точки).

Резюме

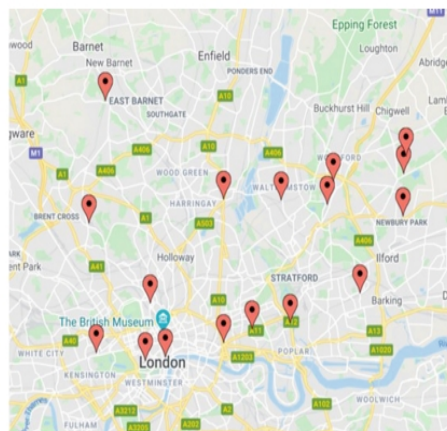
В доклада се разглежда приложението на информационните, компютърните и комуникационните технологии по дисциплината „Физична и Колонидна Химия“. Направен е преглед на опита в създаването на такива курсове по този учебен предмет в българските университети. Представен е електронен курс, създаден в системата за електронно обучение на факултет „Техника и технологии“ – Ямбол, България. Съобразени са изискванията на пазара на труда при създаването на учебната програма по тази дисциплина.

Abstract

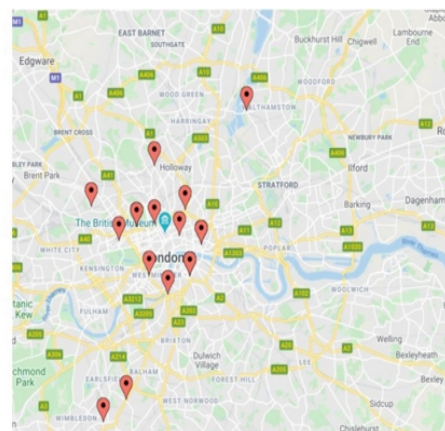
The report deals with the application of information, computer and communication technologies in the course “Physical and Colloid Chemistry”. A review was made of experience in the creation of such courses for this academic subject in Bulgarian universities. Presented is an electronic course, created in the system for e-learning of faculty “Technics and Technologies” – Yambol, Bulgaria. Considered are the requirements of the labor market in the creation of curriculum for this discipline.

София, 08.09.2025 г.

/М.Желязкова/

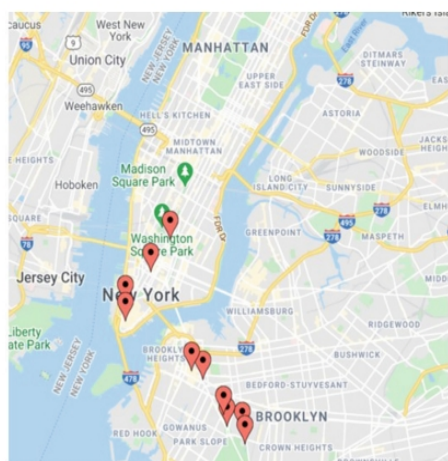


A) Areas with low relative risk ($RR < 1$) for *Staphylococcus aureus* resistant bacteria

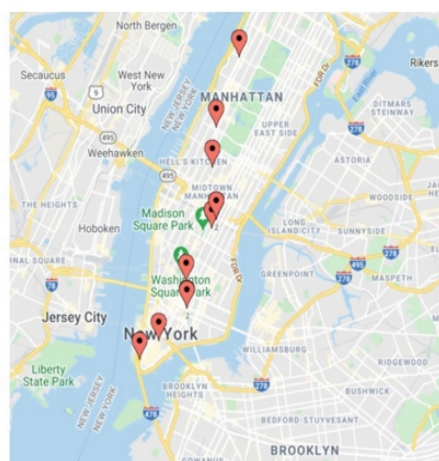


B) Areas with high relative risk ($RR > 1$) for *Staphylococcus aureus* resistant bacteria

Рис. 1: An interactive map of the sampling locations in London with estimated low (A) and high (B) relative risk of *Staphylococcus aureus* AMR.



A) Areas with low relative risk ($RR < 1$) for *Salmonella enterica* resistant bacteria



B) Areas with high relative risk ($RR > 1$) for *Salmonella enterica* resistant bacteria

Рис. 2: An interactive map of the sampling locations in New York City with estimated low (A) and high (B) relative risk of *Salmonella enterica* AMR.

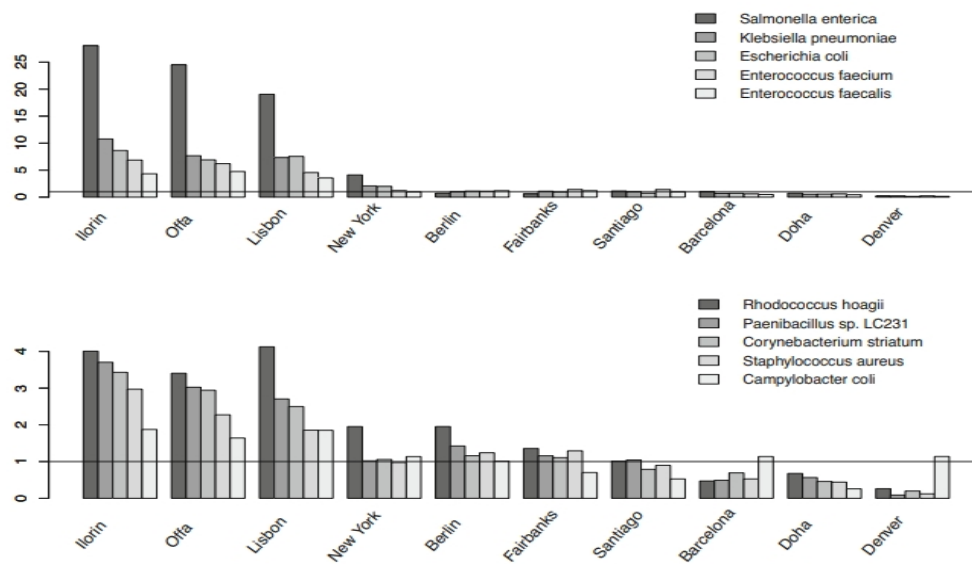


Рис. 3: Barplot of relative risk estimates of top ten abundant antimicrobial resistant species using CSD16 samples. The expected relative risk is the median across the samples for each taxon

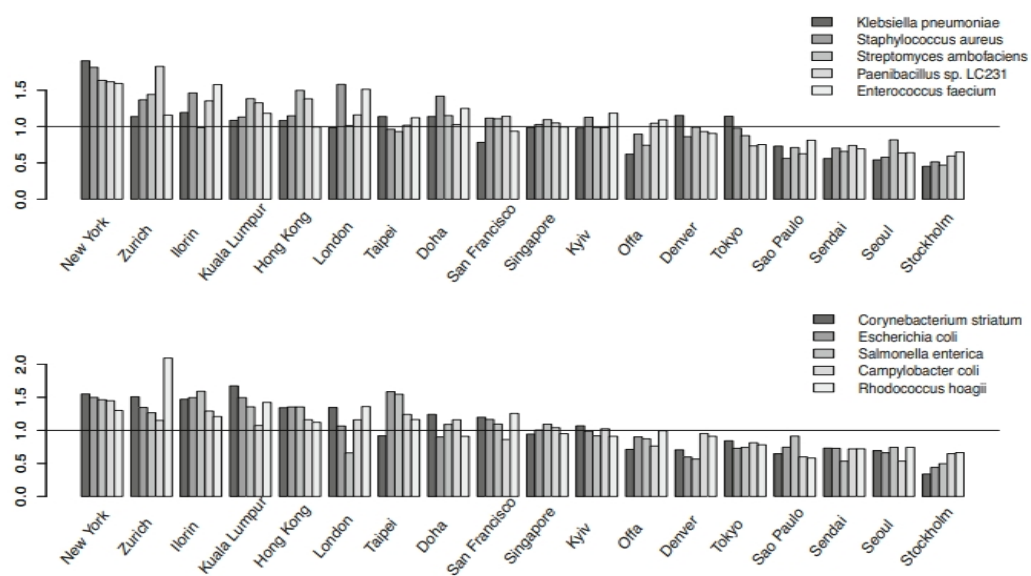


Рис. 4: Barplot of relative risk estimates of top ten abundant antimicrobial resistant species using CSD17 samples. The expected relative risk is the median across the samples for each taxon.