Математички факултет
Универзитет у Београду

Семинарски рад

У оквиру курса Истраживање података 2

Тема:

Истраживање образаца у патогеним острвима секвенци *Escherichia coli* и *Helicobacter pylori*

Професор:
Ненад Митић

Студенти:
Велько Продан, 163/2019
Маја Миленковић, 160/2019

Мај 2024

Sadržaj

1 Увод **3**
1.1 Патогена острва 3
1.2 Ешерихија коли 3
1.3 Хеликобактер пилори 3

2 Алгоритми истраживања образаца **4**
2.1 TF-IDF 4
2.2 NOSEP 4

3 Припрема података **6**

1 Увод

1.1 Патогена острва

Патогена острва (*Pathogenicity island* - PAI) су посебни генетски елементи присутни на хромозомима великог броја бактеријских патогена. PAI кодирају различите факторе вируленције и обично су одсутни код непатогених сојева исте или блиско сродних врста. Сматрају се подкласом геномских острва која се стичу хоризонталним преносом гена путем трансдукције, коњугације и трансформације, и пружају "квантне скокове"у микробној еволуцији, што доприноси способности микроорганизама да еволуирају [2].

Једна бактеријска врста може имати више од једног PAI. PAI су кластери гена уграђени у геном патогених организама, хромозомално или екстрахромозомално. Као тип мобилног генетичког елемента, PAI могу варирати од 10-200 кб. Они носе гене који им омогућавају да производе различите факторе вируленције, укључујући адхезине, токсине, системе секреције и друге елементе који им помажу да се вежу за ћелије домаћина, избегну имуни систем и изазову болест [6]. Подаци засновани на бројним секвенцираним бактеријским геномима показују да су PAI присутни у широком спектру грам-позитивних и грам-негативних бактеријских патогена људи, животиња и биљака. Недавна истраживања усмерена на PAI довела су до идентификације многих нових фактора вируленције које ове врсте користе током инфекције својих домаћина [2].

1.2 Ешерихија коли

Ешерихија коли (Escherichia coli) је разноврсна бактеријска врста која обухвата како безопасне коменсалне сојеве, тако и патогене сојеве који се налазе у гастроинтестиналном тракту људи и топлокрвних животиња. Патогеност Ешерихије коли, која се постиже хоризонталним трансфером гена који одређују факторе вируленције, омогућава овој бактерији да постане врло разноврстан и адаптибилан патоген одговоран за цревне или ванцревне болести код људи и животиња. Сходно томе, сојеви E. коли могу се класификовати у три групе: коменсалне/пробиотске сојеве, интестинално патогене сојеве и екстраинтестинално патогене сојеве.

Растућа количина информација о ДНК секвенцама, генерисаних у "ери геномике", помогла је у повећању разумевања фактора и механизама укључених у диверзификацију ове бактеријске врсте [1].

1.3 Хеликобактер пилори

Хеликобактер пилори је грам-негативни патоген спиралног облика који колонизује антрум и корпус желуца. У последњој деценији, идентификовани су бројни фактори вируленције. Ови елементи омогућавају бактерији да преживи у изузетно киселој средини гастроинтестиналног тракта, доспе до неутралније средине слузног слоја и одупре се имунолошком одговору човека, што резултира перзистенцијом [4].

Сојеви Хеликобактер пилори показују висок степен генетске хетерогености због геномских преуређења, тачкастих мутација, убацивања и/или брисања гена. Генетички јединствене варијанте једног соја присутне су у желуцима сваког човека, а

генетски састав ових популација може се мењати током времена. Ова адаптабилност доприноси и њеној високој заразности [5].

Већина инфекција се јавља у детињству, а само мали проценат инфекција напредује до тежих стања. Инфекција овом бактеријом може изазвати различите гастроинтестиналне проблеме, укључујући хронични гастритис, чир на дванаестопалачном цреву, па чак и рак. Хеликобактер пилори је веома заразна бактерија [4].

2 Алгоритми истраживања образаца

2.1 TF-IDF

TF-IDF[3] (Term Frequency Inverse Document Frequency) алгоритам је статистички метод који се користи за процену важности речи за документ или категорију у скупу датотека или корпусу.

Главна идеја је да ако се нека реч или фраза често појављује у чланку, а ретко се налази у другим чланцима, сматра се да реч или фраза има добру способност разликовања класе и погодна је за класификацију. То је најчешће коришћена функција за израчунавање тежине речи у тренутном векторском моделу простора. Углавном се састоји од два дела, а то су учесталост речи и инверзна учесталост текста. Учесталост речи се односи на број појављивања дате речи у датотеци. Инверзна учесталост датотеке представља меру опште важности речи. Инверзна учесталост речи се дели укупним бројем докумената, који се дели бројем докумената који садрже тај термин, а затим се логаритмује резултат количника. Формуле за учесталост речи (TF) и инверзну учесталост текста (IDF) су следеће:

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$n_{i,j}$ је број појављивања речи ti у датотеци d_j, $\sum_k n_{k,j}$ је збир појављивања свих речи у датотеци d_j.

$$\mathrm{idf}_i = \log\left(\frac{|D|}{|\{j : t_i \in d_j\}|}\right)$$

$|D|$ је укупан број датотека у корпусу, $|\{j : t_i \in d_j\}|$ је број докумената који садрже реч t_i. Ако речи нема у корпусу, то ће довести до деобе са нулом. Зато се уопштено користи $1+|\{j : t_i \in d_j\}|$.

$$\mathrm{tfidf}_{i,j} = \frac{\mathrm{tf}_{i,j} \times \mathrm{idf}_i}{\sqrt{\sum_{t_i \in d_j} [\mathrm{tf}_{i,j} \times \mathrm{idf}_i]^2}}$$

$\mathrm{tfidf}_{i,j}$ је тежина речи t_i. Може се видети да висока учесталост речи у одређеној датотеци и ниска учесталост датотеке речи у целом скупу датотека могу генерисати високу TF-IDF вредност.

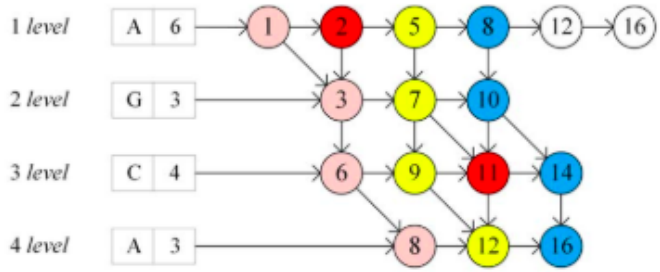
2.2 NOSEP

NOSEP[7] (Nonoverlapping Sequence Pattern Mining With Gap Constraints) представља исцрпан алгоритам за рударење образаца секвенци који користи Nettree

структуру података. Ова структура омогућава прецизно израчунавање учесталости појављивања одређеног обрасца унутар дате секвенце, узимајући у обзир услов непреклапања. Приликом рударења образаца секвенци, два главна фактора која утичу на перформансе су израчунавање подршке и смањење простора кандидата образаца.

За израчунавање подршке, користи се NETGAP алгоритам. NETGAP треба да разликује знакове у секвенци који се могу поново користити за упаривање образаца, да би се испунио услов непреклапања. Овај изазов се односи на немогућност једноставне замене одговарајућих знакова у секвенци знаком "X", јер би то спречило проналажење будућих непреклапајућих појава. Стога, NETGAP мора ефикасно разликовати знакове који се могу поново користити.

Да би се превазишао наведени проблем и имплементирао алгоритам без претраге уназад, предлаже се коришћење Nettree структуре података. Nettree се конструише на основу обрасца и секвенце, а затим се итеративно проналазе минималне путање које представљају појаве обрасца. После сваке итерације, минимална путања и неважећи чворови се уклањају из дрвета. Овај процес се понавља док дрво не постане празно, гарантујући потпуно проналажење свих појава обрасца.



Са друге стране, NOSEP алгоритам користи приступ раста образаца за смањење простора кандидата образаца, који се показао ефикаснијим од приступа претраге у ширину и дубину. NOSEP поседује Apriori својство, што имплицира да ако подобразац није фреквентан, његов надобразац такође не може бити фреквентан. Захваљујући овом својству, простор кандидата образаца се ефикасно редукује.

NOSEP започиње рад проналажењем свих честих узорака дужине 1, а затим итеративно генерише кандидатске узорке растуће дужине комбиновањем честих узорака краћих дужина. За сваки кандидатски узорак, његова подршка се израчунава коришћењем NETGAP алгоритма и Nettree структуре података. Кандидати чија подршка задовољава минимални праг се додају у скуп честих узорака за ту дужину. Овај процес се наставља док се не може пронаћи више честих узорака. На крају, NOSEP враћа унију свих пронађених честих узорака свих дужина, ефикасно рударећи комплетан скуп честих секвенцијалних образаца.

3 Припрема података

Датотеке са геномским секвенцама, и информације о локацијама острва у тим секвенцама, прикупљени су са веб стране *paidb.re.kr*. Геномске секвенце су чуване у FASTA формату, док су индекси почетака и крајева острва чувани у датотекама са одговарајућим суфиксима: *.pai*, *.cpai*, и *.npai*.

FASTA датотеке су учитаване библиотеком Biopython, и секвенце су заједно са индексима острва чуване у Python речнику. За потребе TF-IDF алгоритма, секвенце су конвертоване у ниске.

NOSEP алгоритам је покретан помоћу библиотеке SPMF, за коју је било потребно кодирати секвенце у одређени формат. Улазна секвенца симбола се прво кодира у низ позитивних целих бројева, где сваки број представља одређени симбол. Поред тога, користе се и специјални симболи за раздвајање карактера (-1) и означавање краја секвенце.

На пример, ако имамо следећу секвенцу:

$$A\,A\,G\,T\,A\,C\,G\,A\,C\,G\,C\,A\,T\,C\,T\,A$$

где су симболи кодирани као:

$$1 = A,\ 3 = C,\ 7 = G,\ 20 = T$$

добија се следећа кодирана секвенца:

$$1\,{-1}\,1\,{-1}\,7\,{-1}\,20\,{-1}\,1\,{-1}\,3\,{-1}\,7\,{-1}\,1\,{-1}\,3\,{-1}\,7\,{-1}\,3\,{-1}\,1\,{-1}\,20\,{-1}\,3\,{-1}\,20\,{-1}\,1\,{-1}\,{-2}$$

Literatura

[1] Mickaël Desvaux, Guillaume Dalmasso, Racha Beyrouthy, Nicolas Barnich, Julien Delmas, and Richard Bonnet. Pathogenicity factors of genomic islands in intestinal and extraintestinal escherichia coli. *Front Microbiol*, 11:2065, September 2020.

[2] Ohad Gal-Mor and B Brett Finlay. Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol*, 8(11):1707–1719, August 2006.

[3] Cai-Zhi Liu, Yan-Xiu Sheng, Zhi-Qiang Wei, and Yong-Quan Yang. Research of text classification based on improved TF-IDF algorithm. In *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, pages 218–222, 2018.

[4] Tilahun M, Gedefie A, Belayhun C, Sahle Z, and Abera A. Helicobacter pylori Pathogenicity Islands and Giardia lamblia Cysteine Proteases in Role of Coinfection and Pathogenesis. 2022.

[5] Jennifer M Noto and Richard M Peek, Jr. The helicobacter pylori cag pathogenicity island. *Methods Mol Biol*, 921:41–50, 2012.

[6] Herbert Schmidt and Michael Hensel. Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev*, 17(1):14–56, January 2004.

[7] Youxi Wu, Yao Tong, Xingquan Zhu, and Xindong Wu. Nosep: Nonoverlapping sequence pattern mining with gap constraints. *IEEE Transactions on Cybernetics*, 48(10):2809–2822, 2018.