Raport privind deplasarea în scopul formării profesionale la
Centrul de Biologie al Academiei de Științe din České Budějovice, Republica Cehă.

1. Solicitant (solicitanți) / echipa de cercetare: Dr. Tiberiu SZOKE-NAGY (E4 – Tehnologii Moleculare şi Biomoleculare)

2. Tipul acțiunii: Formare Profesională (FP)

3. Destinația / tematica / durata:

- **Destinația:** Laboratorul de Ecologie Microbiană și Evolutivă, Departamentul de Ecologie Acvatică Microbiană, Institutul de Hidrobiologie, Centrul de Biologie al CAS České Budějovice, Cehia.
- **Tematica**: Noi tehnici de bioinformatică avansată și tehnologii de secvențiere de nouă generație pentru studierea interacțiunii microorganismelor din ecosisteme acvatice.
- **Durata**: 25 Septembrie 2020 – 1 Noiembrie 2020 (37 zile)

Metagenomica se poate defini la modul general ca parte a științelor biologice care se ocupă cu studierea genomurilor aparținând unei comunități biologice dintr-un anumit habitat. Prin genom se înțelege totalitatea materialului genetic (genelor și a informațiilor ereditare) aparținând unui organism (fie el unicelular precum bacteriile sau multicelular) sau unei entități biologice (fagi și virusuri). Altfel spus metagenomica este tehnica de recuperare a genomurilor microbiene (și nu numai) direct din probele de mediu sau clinice, indiferent de natura probei și abundența organismelor.

Analizele metagenomice aplicate probelor de mediu (Fig. 1) presupun explorarea întregii compoziții genetice a comunităților prezente în proba colectată prin: i) izolarea ADN total (metagenomică) sau ARN total (metatranscriptomică) din proba analizată și secvențierea acestuia folosind metode NGS (Next-Generation Sequencing); și ii) analiza bioinformatică folosind diverse aplicații sau algoritmi specifici pentru identificarea genelor (adnotare), asamblarea genomurilor și reconstrucția metabolică a organismelor prezente în proba analizată.

NGS cunoscută și sub denumirea de secvențiere de mare capacitate (en. high-throughput sequencing) este termenul general folosit pentru a descrie o serie de tehnologii moderne de secvențiere. Primul proiect de secvențiere completă a unui genom a fost proiectul genomului uman. In secvențierea genomului uman s-a folosit tehnologia de secvențiere cunoscută sub denumirea de secvențiere Sanger (secvențiere de primă generație). Datorită acestei tehnologii care prezintă o serie de dezavantaje și limitări, întreg genomul uman a fost secvențiat în 13 ani și costul total al acestuia a fost de aprox. 3 miliarde de dolari, fiind finalizat în anul 2003. Spre

deosebire de secvențierea de primă generație, folosind NGS un genom uman ar putea fi secvențiat în mai puțin de o zi și la costuri infime de câteva mii de dolari. Principalele tehnologii de secvențiere NGS fără a specifica caracteristicile și deosebirile acestora sunt: Illumina (Solexa), Roche 454 (pirosecvențiere), Ion Torrent și Nanopore. Principalele avantaje ale tehnologiilor NGS sunt: i) nu este necesară cunoașterea *a priori* a structurii genomurilor; ii) oferă o rezoluție foarte bună chiar și de un singur nucleotid, ceea ce face posibilă detectarea genelor înrudite, transcripturilor cu splicing alternativ, a variantelor de gene alelice și a polimorfismului uninucleotidic; iii) necesită o cantitate mai mică de ADN/ARN (de ordinul nanogramelor); și iv) gradul sporit de reproductibilitate a rezultatelor.

Un studiu metagenomic poate avea două abordări diferite și anume: i) secvențierea ampliconică cunoscută sub denumirea de *metabarcoding* ce implică o etapă preliminară de amplificare PCR a unei gene marker de interes, abordare utilizată când se dorește identificarea taxonomică a comunității din proba analizată; și ii) secvențierea *shotgun* ce implică fragmentarea genomurilor și secvențierea ADN/ARN total din proba analizată, avantajul secvențierii *shotgun* spre deosebire de secvențierea ampliconică este că prin această metodă se pot obține și secvențele genelor marker care ulterior pot fi extrase și utilizate pentru identificarea taxonomică a comunității analizate.

Stagiul de formare profesională s-a realizat la Institutul de Hidrobiologie al CAS din České Budějovice, Republica Cehă și a fost axat pe trei aspecte ce vor fi detaliate în cele ce urmează și anume: i) familiarizarea în ceea ce privește lucrul cu seturi mari de date (zeci de milioane de secvențe); ii) asamblarea *de novo* a unui genom de *Escherichia coli* folosind secvențe disponibile în bazele de date rezultate în urma secvențierii NGS; și iii) identificarea comunității bacteriene dintr-un metagenom obținut din apa unui lac din Republica Cehă.

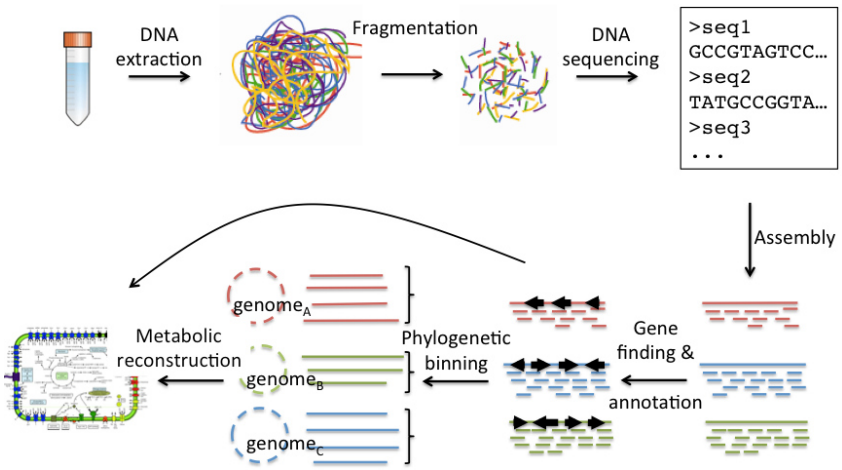

Fig. 1. Reprezentarea schematică a conceptului de metagenomică și prezentarea principalelor etape ale unui studiu metagenomic pentru reconstrucția căilor metabolice din proba analizată. (Sursa http://env gen.github.io/ metagenomics. html#main).

1. **Lucrul cu seturi mari de secvențe** (sute de milioane de secvențe) presupune utilizarea unor programe/algoritmi care să ne permită conversia fișierelor, identificarea diferitelor caracteristici ale secvențelor, redimensionarea setului de date, extragerea anumitor secvențe pe baza unor informații. Desigur în cele ce urmează vor fi enumerați cei mai utilizați algoritmi/programe folosiți în stagiu.

1.1. **Convertseq** este un program care convertește diferite formate standard de secvențe. Acest soft este foarte util atunci când folosim o succesiune de programe care necesită anumite formate pentru a putea rula diferite analize.

1.2. **Lenseq** este folosit atunci când dorim să aflăm doar lungimea secvențelor dintr-un fisier. Programul funcționează cu o varietate mare de fișiere iar rezultatul este afișat pe 3 coloane, care includ ID-ul secvenței, lungimea și descrierea acesteia.

1.3. **GCseq** este similar cu lenseq dar spre deosebire de acesta, gcseq oferă ca și rezultat conținutul de GC (Guanină-Citozină) dint-o secvență.

1.4. **Seqstat** oferă de asemenea informații despre lungimea și tipul secvențelor dintr-un fișier. Fișierul rezultat conține următoarele informații: Formatul fișierului, Tipul de secvență (Proteine, ADN sau ARN), Numărul de secvențe, Numărul total de resturi de nucleotide sau aminoacizi, Lungimea celei mai scurte și celei mai lungi secvențe și Lungimea medie a secvențelor.

1.5. **Reformat** utilizat când dorim să redimensionăm setul de date, de asemenea permite o prelevare randomică a secvențelor dintr-un fișier inițial.

1.6. **faSomeRecords** este utilizat pentru extragerea de secvențe din fișiere conținând milioane de secvențe pe baza unor criterii stabilite inițial.

1.7. **Foxhound2** este similar cu **faSomeRecords** dar avantajele față de acesta sunt reprezentate de rapiditatea cu care se realizează extragerea informațiilor din fișiere precum și faptul că suportă mai multe formate de secvențe.

1.8. Comenzile **head** și **tail** sunt utilizate când dorim să vedem începutul sau finalul unui fișier. Este utilizat frecvent pentru a ne asigura că fișierele rezultate după fiecare etapă au fost generate corect.

1.9. Filtrarea secvențelor se poate face cu ajutorul softurior **daffy**, **sieve**, **keeplong**, **keepshort**, **lenfilter** sau **gcfilter**. Cu ajutorul acestora se pot extrage secvențe în funcție de diferiți parametri precum: lungimea acestora sau conținutul de GC.

1.10. **Fa2aacomp** permite aflarea compoziției de aminoacizi dintr-o secvență, softul folosește formate fasta iar rezultatul oferit este tabelar și cuprinde: ID-ul secvenței, descrierea și lungimea acesteia precum și compoziția celor 20 de aminoacizi.

2. **Asamblarea *de novo* a genomului la *E. coli***

Asamblarea genomului la *Escherichia coli* K12 sa realizat folosind softurile Spades vers. 3.14.1 varianta pentru Linux (link: https://cab.spbu.ru/software/spades/) sau Megahit vers. 1.2.9 varianta pentru Linux (link: https://github.com/voutcn/megahit). Aceste două softuri sunt gratuite și pot fi descărcate și folosite de la linkurile menționate. Evaluarea calității asamblării s-a realizat folosind softul Quast vers. 3.0 (link: https://github.com/ablab/quast).

Pentru început au fost descărcate fișierele fastq ce conțin secvențele rezultate în urma secvențierii NGS. Fișierele forward (ERR022075_1.fastq.gz) și reverse (ERR022075_2.fastq.gz) pot fi descărcate de la linkul următor: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR022/ERR022075/. Ca si referință s-a utilizat genomul de la *E. coli* K12 MG1655 disponibil în NCBI și având codul de acces U0096.

După ce au fost descărcate fișierele conținând secvențele s-a folosit aplicația seqstat pentru a vedea conținutul fișierelor, astfel fișierele ERR022075_1 și ERR022075_2 conțin un număr de 22.720.100 secvențe, 2.272.010.000 nucleotide și o medie de 100 nucleotide per secvență (read). Fișierul de referință conține o singură secvență având lungimea de 4.641.652 nucleotide, aceasta fiind și lungimea genomului la tulpina *E. coli* K12 MG1655.

Următoarea etapă în procesul de asamblare este eliminarea ambiguităților și a regiunilor cu calitate slabă (Quality trimming). Controlul calității se poate realiza cu ajutorul softului seqtk trimfq sau bbduk.sh din bbmap (http://seqanswers.com/forums/showthread.php?t=42776). Fișierele rezultate în urma acestei etape au următoarele caracteristici: i) ERR022075_1 prezintă un număr de 22.502.287 secvențe, 2.163.429.273 nucleotide și o lungime medie a secvențelor de 96.1 nucleotide; iar ii) ERR022075_2 prezintă un număr de 22.502.287 secvențe, 2.219.276.426 nucleotide și o lungime medie a secvențelor de 98.6 nucleotide.

După realizarea controlului calității, fișierele rezultate au fost folosite pentru asamblarea genomului. Asamblarea genomului s-a realizat comparativ folosind 2 abordări diferite: i) prima dată s-a asamblat genomul folosind k-meri unici de lungimi diferite și anume 29, 59, 91, 93 și 95 de nucleotide (Fig. 2) respectiv o succesiune a acestora și programul identifică cea mai bună variantă (multi); și ii) a doua-a a vizat redimensionarea setului inițial de secvențe astfel încât să avem o acoperire a genomului de 10 X, 100 X, 500 X și toate secvențele (Fig. 3).

După cum se poate observa din Fig. 2, cele mai bune rezultate au fost obținute cu Megahit și Spades în cazul în care s-a folosit o succesiune de K-meri (megahit_multi.fasta – negru și spades_multi.fasta – gri), respectiv când setul de date folosit are o acoperire a genomului de minim 100 X (Fig. 3). Obiectivul unei asamblări reușite este să se ajungă în faza de platou folosind cât mai puține contiguri posibile, ideal ar fi 1 (ceea ce înseamnă că genomul asamblat

este reprezentat de un singur contig). Prin redimensionarea setului inițial de date s-au putut asambla genomuri, acestea având 95,835% și 98,295% nucleotide comparativ cu genomul de referință.

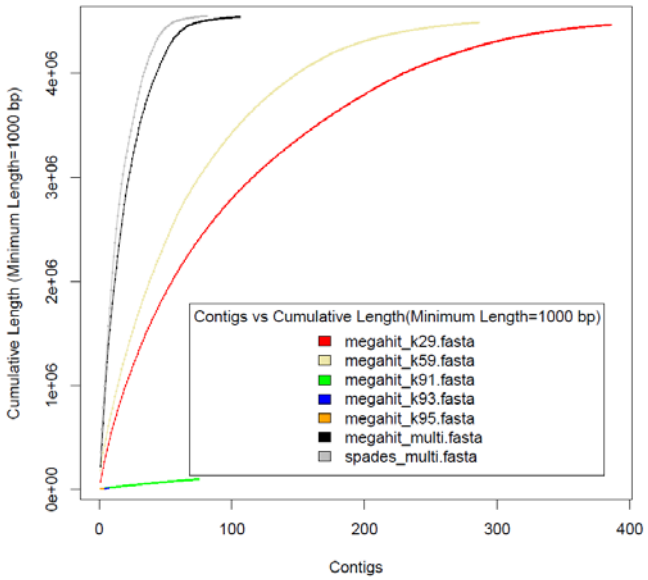

Fig. 2. Asamblarea genomului la *Escherichia coli* K12 folosind k-meri unici și o combinație a acestora. Contigurile cu lungime mai mică de 1000 bp au fost eliminate din grafic.

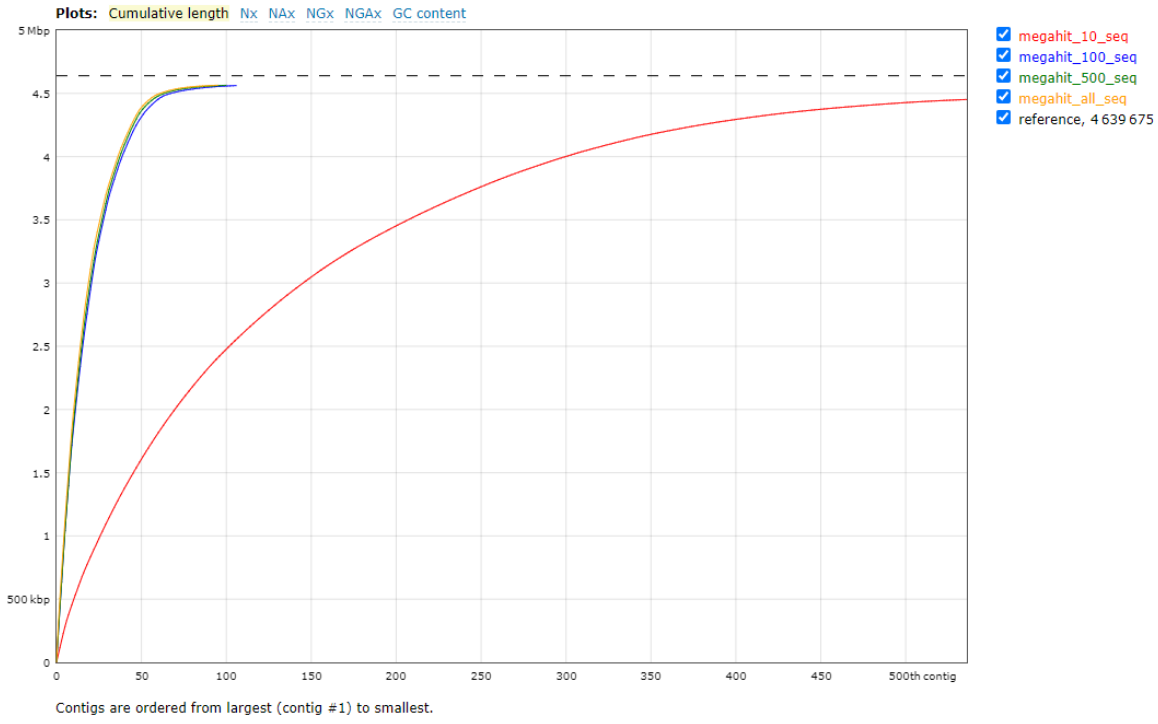

Fig. 3. Asamblarea genomului la *Escherichia coli* K12 folosind megahit rezultat în urma redimensionării setului inițial de secvențe la o acoperire de 10 X, 100 X, 500 X și toate secvențele.

3. **Identificarea comunitatii bacteriene dintr-o probă de apă**

Identificarea secvențelor de ARNr 16S s-a realizat folosind un metagenom recuperate din apă unui lac din Cehia. Pentru comparație setul inițial de secvențe a fost redimensionat la 10, 50 de milioane secvențe respectiv tot setul de date.

Înainte de redimensionarea setului de secvențe, acestea au fost supuse unui control de calitate după cum am menționat anterior. Ca și referință s-a utilizat baza de date Silva SSURef_NR99 disponibilă pentru a putea fi descărcată și utilizată la link-ul următor: https://www.arb-silva.de/no_cache/download/archive/current/Exports/.

Căutarea secvențelor de 16S dintr-un metagenom s-a realizat folosind softul mmseq2, disponibil aici: https://github.com/soedinglab/mmseqs2/wiki. Căutarea secvențelor cu mmseq2 este similară unei căutări blastn, dar implementarea mmseq2 oferă avantajul obținerii mai rapide a rezultatelor. Linia de comandă pentru mmseq2 folosită pentru identificarea secvențelor de 16S este: ./mmseq2 tipul_căutării (easy-search) fișier_inițial.fasta baza_de_date.fasta rezultat.txt Foldertemporar –optiuni. Rezultatul obținut este sub forma unui fișier având formatul m8, delimitat prin tab.

Din fisierul rezultat în urma analizei cu mmseq2 se extrag doar secvențele (readurile) care conțin secvențe referitoare la 16S cu ajutorul programelor faSomeRecords sau foxhound2. În urma acestei procesări am obținut un număr de aproximativ 13.000 secvențe potențiale de 16S dar care conțin și secvențele de 18S (eucariote) respectiv 16S de la Archaea. Gruparea acestora se realizează folosind softul ssu-align (link: http://eddylab.org/software/ssu-align/) rezultând fișiere de tipul archaea.fa, bacteria.fa și eukarya.fa. Fișierele rezultate sunt folosite împreună cu silva-sloth pentru identificarea taxonomică a secvențelor extrase din metagenom. Reprezentarea grafică a principalelor grupe taxonomice este redată în Fig. 4.

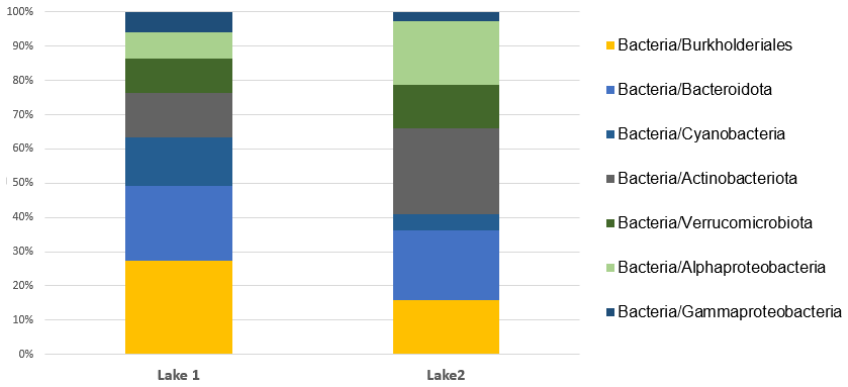

Fig. 4. Principalele grupe taxonomice de bacterii prezente în metagenomurile analizate.

Informațiile și cunoștințelor dobândite în urma deplasării vor fi utilizate pentru realizarea unui model experimental/procedeu despre asamblarea *de novo* a genomurilor bacteriene unde vor fi detaliați pașii ce trebuie parcurși pentru asamblarea genomurilor și pentru dezvoltarea de noi direcții de cercetare în cadrul institutului în concordanță cu strategia actuală și viitoare de dezvoltare a INCDTIM.

SZOKE-NAGY Tiberiu 6 Octombrie 2020