DOI:10.15774/PPKE.ITK.2014.006

A bakteriális kommunikáció és kooperáció génjeinek elhelyezkedése ismert genomokban.

Az AHL szabályzórendszer génjei.

Pázmány Péter Katolikus Egyetem

Információs Technológiai és Bionikai Kar

Multidiszciplináris Műszaki és Természettudományi Doktori Iskola

Gelencsér Zsolt

Konzulens: Prof. Pongor Sándor

2014

DOI:10.15774/PPKE.ITK.2014.006

Köszönetnyilvánítás

Elsősorban szeretnék köszönetet mondani témavezetőmnek, Dr. Pongor Sándor professzornak, aki segítette és irányította a tanulmányaimat. Nélküle ez a munka nem jöhetett volna létre. Továbbá szeretném megköszönni a biológiai és informatikai kérdésekben való segítségnyújtást és tanácsadást Kumari Sonal Choudharynak és Sanjarbek Hudaiberdievnek. Köszönettel tartozom Dr. Vittorio Venturi professzornak és csoportjának, akik nem csak elláttak a munkámhoz szükséges adattokkal, hanem az általam kinyert információk ellenőrzésében is segítettek.

Szeretném megköszönni a szakmai segítségnyújtás a PPKE ITK doktoranduszainak, kiváltképp Bihary Dórának és Ligeti Balázsnak. Továbbá köszönöm a segítségét Galbáts Borisznak és Erdei Áronnak, akik szakdolgozattukkal segítették a munkám előrehaladását.

Hálás vagyok a PPKE ITK doktori iskolájának és vezetőinek Dr. Roska Tamás és Dr. Szolgay Péter professzoroknak, hogy lehetőségek biztosítottak a munkám zavartalan elvégzéséhez, valamint a trieszti ICGEB kutatóintézetének, hogy részvehettem a kurzusain, melyek elmélyítették tudásomat a bioinformatika szakterületein.

DOI:10.15774/PPKE.ITK.2014.006

Kivonat

Habár a baktériumok egyszerű mikrobiális élőlények, mégis képesek közösségeket alkotni, amelyek bonyolult viselkedésminta létrehozására is képesek. Ehhez alapvető feltétel, hogy tudjanak egymással kommunikálni és kooperálni. Ennek a folyamatnak pontos megismerése azért fontos, mert ezek a bakteriális közösségek jelentős szerepet játszanak az élettudomány legtöbb területén, mint például fertőzések terjedésében, a talaj- és tengeri bioszféra egyensúlyában. A ma talán legjobban ismert kommunikációs mechanizmus az úgy nevezett *quorum sensing*, amelynek során a baktériumok külső kémiai jelanyagokkal érintkeznek egymással. Az egyik jelentős ilyen jelanyag az *acil homoszerin laktón* (AHL).

Munkám célja, hogy egy automatizált genomannotációs eljárás segítségével megkeressem a fent említett kommunikációs eljárás fehérjéit kódoló géneket a ma ismert bakteriális genomokban, és megvizsgáljam, hogy van-e szerepe a gének egymáshoz viszonyított helyzetének és irányának a kommunikáció mechanizmusában. Az annotációs eljárás egy rejtett Markov modell alapú, általános alrendszerkereső algoritmus, így más biológiai rendszerek vizsgálatára is használható. A folyamat során több adathelyesség ellenőrzés is történik, hogy a végső annotáció minél megbízhatóbb legyen.

Munkám során kimutattam, hogy a regulátor/szenzor fehérje (*LuxR*) és a jel-szintetizáló fehérje (*LuxI*) génjei legtöbbször párban állnak, és gyakran egy szabályzó fehérje (*RsaM, RsaL*) helyezkedik el közöttük. Azt is tapasztaltam, hogy ezek a gének csak bizonyos, jól definiálható elrendeződésekben találhatóak meg, így ennek valószínűleg szerepe van a kommunikációs folyamat mechanizmusában. Szekvenciálishasonlóság keresések segítségével sikerült észrevennem azt is, hogy különböző fajok azonos elrendeződésű *quorum sensing* génjei jobban hasonlítanak egymásra, mint az azonos fajban szereplő, különböző elrendeződések.

DOI:10.15774/PPKE.ITK.2014.006

Abstract

Bacteria might be simple, single celled organisms but their social behavior means they can form complex communities and engage in coordinated behaviors. For this it's a fundamental they capable to communicate and cooperate each other. The understand of this mechanism is important, because this bacterial communities play significant role in many part of life science, such as infection spreading, balance of soil- and marine biosphere. One of the most well studied communicational mechanisms is the so-called *quorum sensing* that use external chemical transmitters to deliver information. One of the most important is the acyl-homoserine lactones (AHL).

The purpose of this work is to search the protein-coding genes of the aforementioned communicational process in the today known bacterial genomes via an automata genome annotation algorithm and to analyze the role of the gene orientation and arrangement in the communicational mechanism. The annotation algorithm is a Hidden Markov Model based general subsystem search method, so it is capable to analyze other biological systems. During the process there are more data correction checks to increase the reliability of the result.

I revealed that the regulator/sensor protein (*LuxR*) and the signal-synthase protein (*LuxI*) mostly appear in pair, and often there is a regulator protein between them. I observed that this genes appear only a few conserved arrangements, so it is probably has role in the communicational process. Via sequence similarity search I noticed, that the same *quorum sensing* genes in different species are more similar each other than different arrangements that appears in the same species.

DOI:10.15774/PPKE.ITK.2014.006

Tartalomjegyzék

DOI:10.15774/PPKE.ITK.2014.006

DOI:10.15774/PPKE.ITK.2014.006

Ábrajegyzék

DOI:10.15774/PPKE.ITK.2014.006

Táblázatjegyzék

DOI:10.15774/PPKE.ITK.2014.006

Rövidítésjegyzék

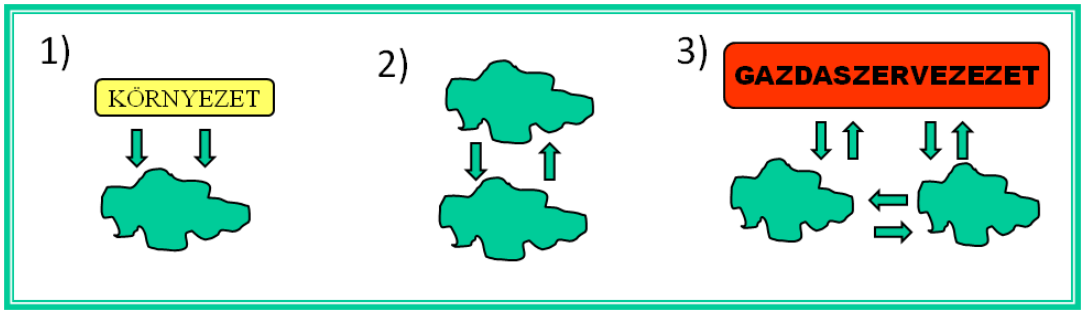
QS	quorum sensing
AHL	acil homoszerin lakton
DNS	dezoxiribonukleinsav
IUPAC	International Union of Pure and Applied Chemistry
HTH	Helix-turn-Helix
PAM	Percent Accepted Mutation
BLOSUM	Blocks Substitution Matrix
HMM	Hidden Markov Model
BLAST	Basic Local Alignment Search Tool
HSP	High-scoring Segment Pair
ORF	Open Reading Frame
GI	GenInfo Identifier
CDS	coding sequence
NGS	New Generation Sequencing
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
COG	Clusters of Orthologous Group
EBI	European Bioinformatics Institutes
EMBL	European Molecular Biology Laboratory
DDBJ	DNA Database of Japan
PDB	Protein Data Bank
RCSB	Research Collaboratory for Structural Bioinformatics
KEGG	The Kyoto Encyclopedia of Genes and Genomes
GO	Gene Ontohology
HTML	HyperText Markup Language
XML	Extensible Markup Language
FTP	File Transport Protocol
CSS	Cascading Style Sheets

DOI:10.15774/PPKE.ITK.2014.006

1. Bakteriális kommunikáció: Quorum Sensing

A mikroorganizmusok, így a baktériumok is képesek arra, hogy csoportosan a többsejtű élőlényekhez hasonló viselkedés mintákat hozzanak létre. Ehhez elengedhetetlen, hogy a baktériumok valamilyen módon információhoz jussanak a populáció többi tagjával kapcsolatban, azaz valamilyen módon kommunikáljanak egymással. Ennek egyik legegyszerűbb módja a kémiai jelanyagok segítségével történő kommunikáció, melyek révén a sejtek információt nyernek a környezetükről, és bizonyos esetekben szabályozzák a populáció sűrűségét is.

A kommunikáció eszközei főként a másodlagos *metabolit*ok sorából kerülnek ki. Az elsődleges *metabolit*ok a sejt fő alkotóelemei: a fehérjék, nukleinsavak, szénhidrátok, illetve ezek alkotóelemei. A másodlagos *metabolit*ok csoportjába soroljuk mindazokat az anyagokat, amelyeket a sejt ezen felül termel. A baktériumok különösen gazdag forrásai az ilyen anyagoknak. A másodlagos *metabolit*ok nagy részét a sejtek kibocsájtják magukból, egyszerűsítve azt mondhatjuk, hogy a baktériumok a másodlagos *metabolit*ok felhőjében mozognak. Ide tartoznak például az antibiotikumok és a jelzőanyagok is. Azt, hogy egy anyag jelnek tekinthető-e, John Maynard Smith nyomán evolúciós szempontok alapján szokták definiálni. [1] Maynard/Smith definíciója szerint egy anyag akkor jel, ha mind a kibocsátó, mind a felfogó szervezetben evolúciósan rögzült mechanizmusok alakultak ki a kommunikációra. Egy baktérium természetesen reagálhat bármely anyagra, de ezeket nem jelnek, hanem kulcsnak, *clue*-nak szokták nevezni. Az anyag nem tekinthető jelnek, ha nem azzal a céllal termelték, hogy az adott baktérium reagáljon rá. A bakteriális kommunikáció típusait, és a tanulmányozási módszereiket az 1.1. ábra foglalja össze.



1.1. ábra A bakteriális kommunikáció 3 főtípusa

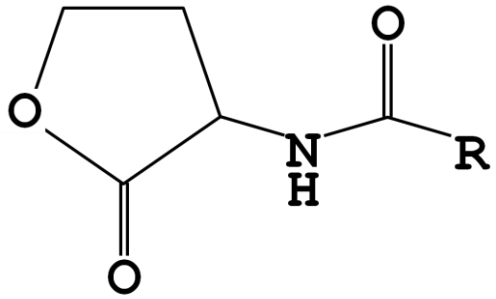
1) A baktériumoknak sok jelérzékelő rendszere van, amellyel a környezet különböző anyagait érzékelik. Például a *kemotaxis*.

2) A baktériumok sokszor a saját maguk illetve a saját fajuk által termelt anyagokra is reagálnak. Például a *quorum sensing* esetében.

3) Egy gazdaszervezet baktériumai egymással és a gazdaszervezettel is kommunikálnak, ezzel létrehozva komplex egységet. Például a bélflóra.

DOI:10.15774/PPKE.ITK.2014.006

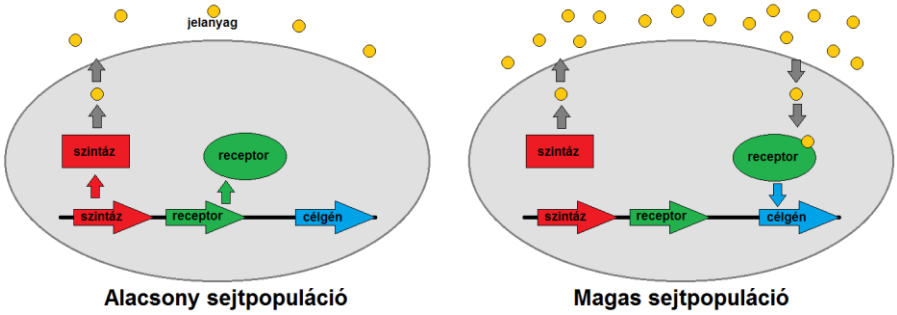
A bakteriális kémiai jelanyagok témájának úttörő tanulmánya volt a *Vibrio fischeri* foszforeszkálás szabályozásának leírása [2]. A *V. fischeri* egy Hawaii-i tintahal faj (*Euprymna scolopes*) fénytermelését szabályozza; a sejtközötti térben jelenlévő jelanyag koncentrációjának egy meghatározott szintje olyan hatás mechanizmust vált ki, ami fény kibocsátással jár. Csak több mint 10 évvel később azonosították a jelanyagot (AHL = *acil homoszerin lakton* [3]) és a szabályzásért felelős fehérje párt: a jelanyag termelő *LuxI* fehérje és a *LuxR* receptor fehérje [4]. Ez a kommunikációs forma a *quorum sensing* nevet a 90-es években kapta [5], amiről mára már tudjuk, hogy az egyik legfontosabb alapja a baktériumok közösségi viselkedésének. [6-8] A mechanizmus alapja a kémiai jelek által kiváltott sejttevékenységek, mint például az osztódás, fertőzési faktorok termelése, vagy túlélést segítő tevékenységek. A különböző baktérium fajok közötti kommunikáció tovább növeli a túlélés esélyét.



1.2. ábra N-acil homoszerin lakton általános kémiai felépítése

A legjobban ismert *quorum sensing* mechanizmus az N-AHL-en (N-acil homoszerin lakton) alapul.(1.2. ábra) A baktérium sejtek folyamatosan figyelik a jelanyagot, hogy ismerjék a környezetüket, és reagálni tudjanak a történt változásokra. Ha a jelanyag elér egy kritikus értéket, akkor a baktérium populáció tagjai egy célgén megváltoztatott kifejtődése által érik el a kívánt viselkedést. Az N-AHL alapú *quorum sensing* két fehérjén alapul, amelyek a *LuxI* és *LuxR* családok tagjai. [6, 9] A *LuxI* családba tartozó fehérjék felelősek az N-AHL szintéziséért. [10] A szintézis után a jelanyag szabadon átjuthat a sejtfalon, így a sejten belül és a sejten kívül is felhalmozódhat a sejtsűrűség függvényében. Ha a populáció kis méretű, a jelanyag szétszóródik, ha a populáció nagy, a jelanyag koncentrációja nő. Egy kritikus koncentráció érték felett az N-AHL kapcsolatba lép a *LuxR* családba tartozó fehérjével [11], ami legtöbb esetben egy olyan komplexet eredményez, ami egy speciális DNS szakaszhoz kötődik a célgén promoter régiójában. [12] Ez a génkifejtődés segítségével hatást fejt ki az élőlény fenotípusára. (1.3. ábra) Gyakran előfordul, hogy a célgének között szerepel *LuxI* családbeli gén is, így létrehozva egy pozitív visszacsatolású hurkot.

DOI:10.15774/PPKE.ITK.2014.006



1.3. ábra A quorum sensing mechanizmusának vázlata

A *quorum sensing* működése alacsony illetve magas jelanyag koncentráció esetén. A jelanyag koncentrációja arányos a sejtpopulációval; az adott területen minél több baktérium található, annál magasabb lesz a jelanyag koncentrációja.

Az előbb leírt folyamat már régóta ismert a biológusok körében. Mint látható, jelen esetben egy pozitívan visszacsatolt rendszerről van szó, ami ebben a formában nem lehet stabil, mert a jelanyag termelés az önindukció hatására végtelenségig nőne. Ezzel ellentétben áll, hogy a természetben és a kísérletekben nem tapasztalható instabilitás. Ez enged következtetni arra, hogy valamilyen szabályzó rendszer egészíti ki ezt a kommunikációs rendszert, ami stabilizálja a folyamatot és megakadályozza, hogy a baktériumok felesleges jelzőanyag termelésre pazarolják az erőforrásaikat.

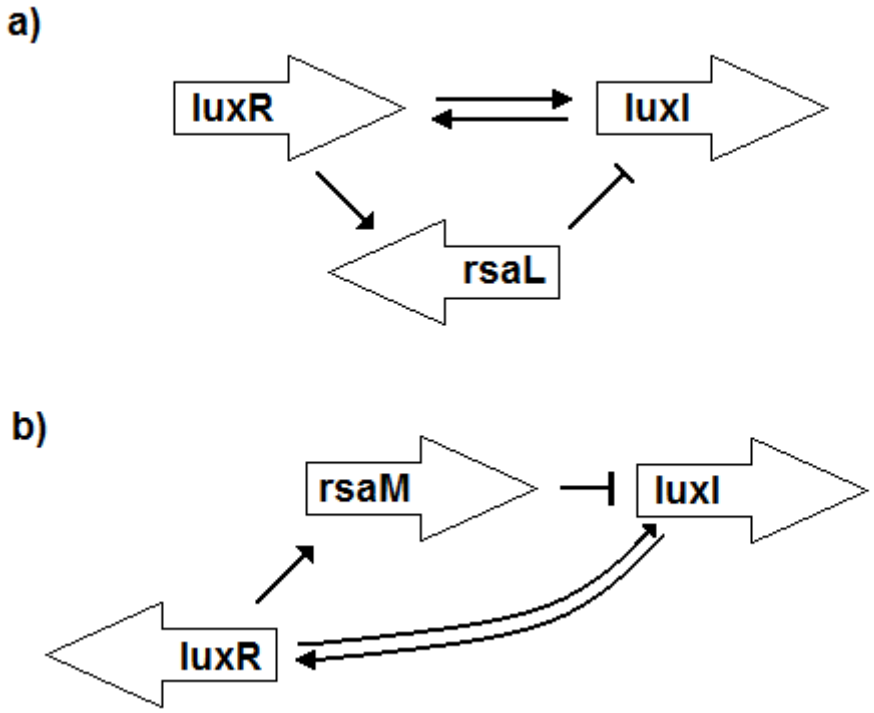
1.1. A quorum sensing szabályzó fehérjéi

Vannak baktériumok, amelyekben a *luxR* és *luxI* géneken kívül egy harmadik, szorosan mellettük elhelyezkedő szabályzó gén is megjelenik. Ezek közül a gének közül kettő fontosat emelnék ki: az *rsaL* és az *rsaM* géneket.[13, 14] Mindkettőre igaz, hogy leggyakrabban *luxR* és *luxI* gén párok közvetlen közelében helyezkednek el, pontosabban a két gén között. Érdekesség, hogy míg az *rsaM* gén csak *quorum sensing*gel rendelkező baktériumban található, *rsaL*-szerű gén más baktérium fajokban is megtalálható, amelynek szerepe egyelőre még nem tisztázott.

A fent említett két szabályozó gén bemutatására a *Pseudomonas fuscovaginae* baktérium a legalkalmasabb, mivel ebben a fajban ismerték fel a működésüket, és azon ritka baktériumok közé tartozik, ami mindkét mechanizmussal rendelkezik.[14]

Ennek a fajnak szintén *quorum sensing* rendszere van: *PfvI*/*PfvR*, *PfsI*/*PfsR*. Az *rsaL* génje a *PfvI*/*PsvR* rendszerben, míg az *rsaM* gén a *PfsI*/*PfsR* rendszerben található. Kísérleti úton bizonyították, hogy az *RsaM* fehérje képes ebben a baktérium fajban a *PfvI*-t és a *PfsI*-t is represszálni. Ezzel szemben az *RsaL* csak a *PfvI* kifejeződését gátolja. A stabilizáló szerepe a két, még kevéssé vizsgált fehérjének azért valósul meg, mert a *PfvI* és a *PfsI* a jel előállításáért felelnek.(1.4. ábra)

DOI:10.15774/PPKE.ITK.2014.006



1.4. ábra A quorum sensing szabályozás negatív visszacsatolása

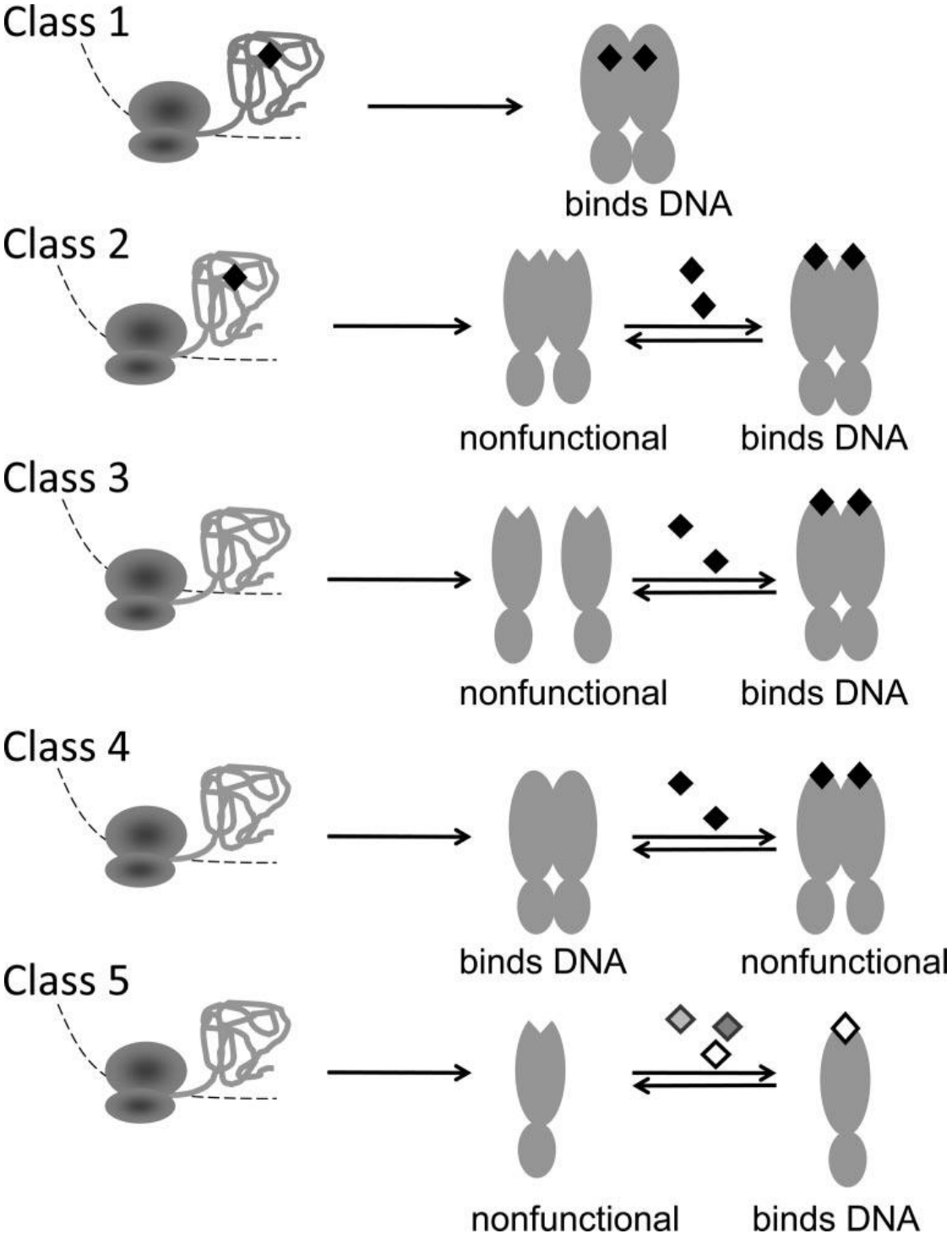
A *quorum sensing* szabályozás *rsaL* és *rsaM* génjeinek sematikus elhelyezkedése. A géneket reprezentáló irányok azt jelzik, hogy a gének melyik DNS szálon vannak. A szabályzó nyilaknál a rendes nyíl a felszabályozás, a kalapácsfejű nyíl a leszabályozás jele. a) Az *rsal* gének elhelyezkedése. b) az *rsaM* gének pozíciója.

1.2. A LuxR fehérje típusai

A *LuxR* fehérje családnak tagjait több alcsoportra is oszthatjuk, attól függően, hogy milyen kölcsönhatásban vannak az AHL molekulával illetve milyen multimerik tulajdonsággal rendelkeznek (1.5. ábra). [15]

1. **csoport:** Tagjai az AHL molekulával cotranszlációsan csatlakozik, és a jelanyag a fehérje struktúrába épül.
2. **csoport:** Az AHL molekula stabilizálja a fehérjét a transzláció alatt, habár maga a jelanyag kötés reverzibilis folyamat. Ilyen receptor például a *LuxR* fehérje.
3. **csoport:** Tagjai stabilak ugyan AHL nélkül is, de a dimerizációhoz és a transzláció aktiválásához szükséges a jelanyag jelenléte.
4. **csoport:** Tagjai az AHL hiányában is képesek a DNS megkötésére, de a jelanyaggal való kapcsolósás deaktiválja a fehérjét.
5. **csoport:** Tagjai nem dimerizálnak és képesek több nemrokon AHL jel felismerésére.

DOI:10.15774/PPKE.ITK.2014.006



1.5. ábra A LuxR fehérje típusainak bemutatása

Az adott LuxR fehérjével rokon AHL molekula fekete rombusszal, míg a nem-rokon jelanyag fehér illetve szürke rombusszal van jelölve az ábrán. Az ábra a *Stevens et al.* cikkből származik. [15]

DOI:10.15774/PPKE.ITK.2014.006

2. Bevezetés

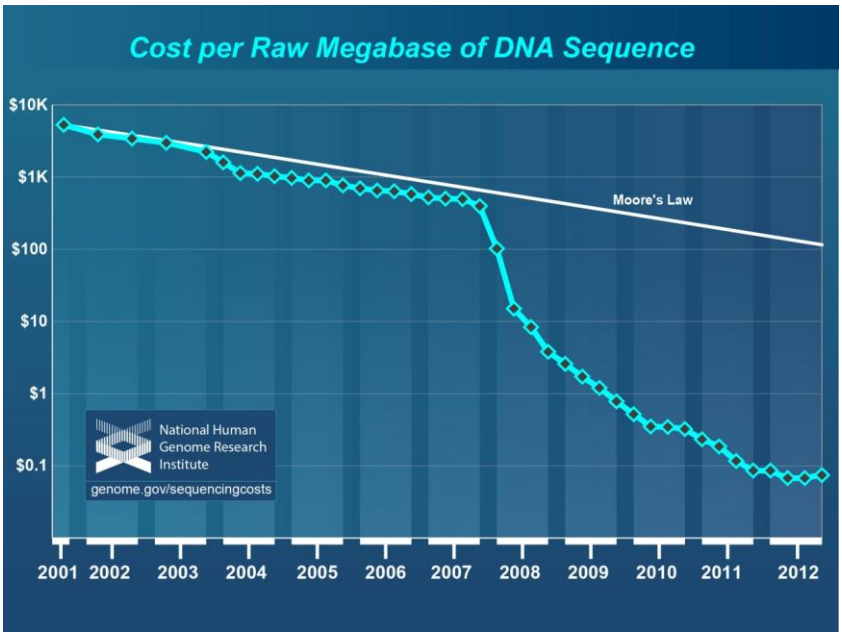
2.1. A DNS szekvenálás

Még 20 év sem telt el a *Haemophilus influenzae* baktérium teljes genomjának szekvenálása óta [16] (ami az első ilyen eredmény volt), ma már a több mint 2100 baktériumon kívül 150 eukarióta teljes genomja is ismert, a folyamatban lévő szekvenálások száma pedig a 15 000-et is meghaladja.[17] Amíg eddig eljutottunk, több lényeges eredmény is született. A legfontosabb mérföldkő azonban minden kétséget kizáróan a 2000 júniusában következett el, amikor az amerikai elnök és a brit miniszterelnök bejelentette a humán genom projekt első fázisának befejezését: feltárták az emberi genomot. Ekkor ugyan még csak nyers „*draft*" formátumban volt kész, de az ezt követő években elkészült a teljes genommá rendezés is.

A szekvenálási módszerek rohamos fejlődésével az új szekvenciák száma is exponenciális mértékben nő. Míg korábban a Sanger és társai által kidolgozott „*chain-termination*"-ön alapuló módszert használták [18], addig mára a legtöbb esetben új generációs (NGS = New Generation Sequencing) szekvenálási eljárásokat használnak.[19] Ezen módszerek legnagyobb előnye, hogy nagyságrendekkel olcsóbbak, mint a régebbi eljárások. Míg 2001-ben egymillió bázis szekvenálásának költsége közel 5300$-ba került, addig 2012 januárjában ez a szolgáltatás már 0,09 $-os áron volt elérhető.[20] (2.1. ábra) A szekvenálás költségének nagyarányú csökkenése a megismert DNS láncok számának nagymértékű növekedését hozta. Az NCBI (National Center for Biotechnology Information) GenBank adatbázisa az egyik legfontosabb elsődleges DNS szekvencia adatbázis, aminek a mérete a 2008-as statisztika szerint már a 100 milliárd bázispárt is elérte, pedig 10 évvel előtte még csak egy-két milliárd bázispárt tartalmazott. (2.2. ábra)

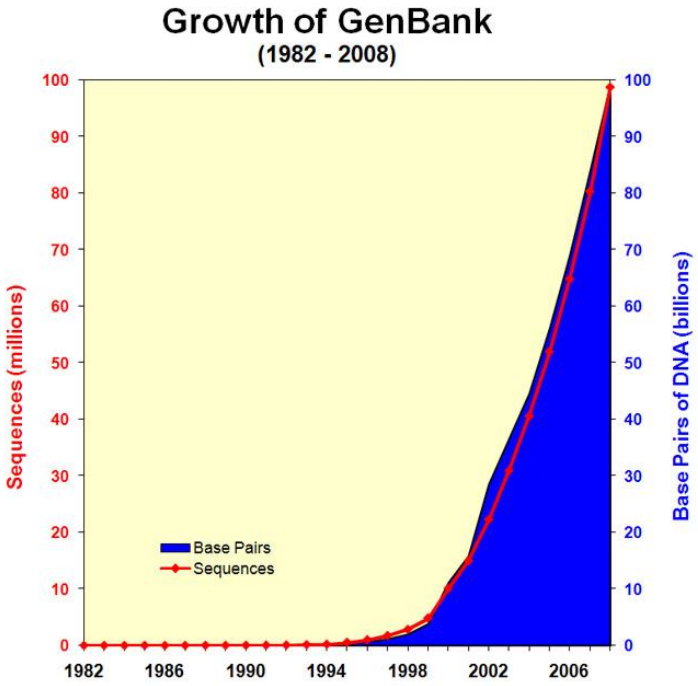
A szekvenciák számának rohamos növekedése mellett fontos feladat a szekvenciák által kódolt működés megismerése is: hiába rendelkezünk például a teljes humán genommal, meg is kell értenünk annak felépítését és funkcióját, különben egy egyszerű szövegsorozat marad biológiai háttér nélkül. Az ismeretlen szekvenciák megismerésének eszköze a genomannotáció.

DOI:10.15774/PPKE.ITK.2014.006



2.1. ábra A szekvenálás költségének változása az elmúlt évtizedben

A grafikon évekre bontva mutatja egy megabázis szekvenálásána költségét összehasonlítva a Moore-törvénnyel (amely a tapasztalati megfigyelésen alapuló technikai fejlődés mértéke). Jól látható, a 2000-es évek elején nagyjából követte az általános tendenciát, de 2007 végétől jelentősen csökkent a költség. [21]



2.2. ábra A GenBank növekedése az elmúlt évtizedekben

A grafikon az NCBI GenBank adatbázisának méretbeli növekedését mutatja a 2008-as évvel bezárólag. Jól látható az elmúlt évtizedben történt robbanásszerű növekedés.[22]

DOI:10.15774/PPKE.ITK.2014.006

2.2. A genomannotáció fogalma

A genomok annotációja informálisan azt a folyamatot jelöli, melynek során a nyers genomiális szekvenciához bónusz információkat adunk hozzá. A MedicinedNet.com definíciója tipikus példája az ilyen informális definícióknak: „*Genome annotation: The process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do. An annotation (irrespective of the context) is a note added by way of explanation or commentary. Once a genome is sequenced, it needs to be annotated to make sense of it.”* A nyers adatokhoz már azok előállításakor adnak némi annotációt (például, hogy milyen szervezetből valók, de legalábbis annak a kísérletnek a számát, ami közben meghatározták), de a komolyabb annotációs munka akkor kezdődik, mikor a nyers adatokat hozzáadják egy adatbázishoz, (főleg egy nyilvános, mások számára hozzáférhető és hivatkozható adatbázishoz). Ezt a munkát általában biológiai és bioinformatikai ismeretekkel rendelkező annotátorok végzik, akik munkájuk során egy sor, különböző adatbázist, bioinformatikai programot használnak, és az eredmények megfogalmazására kötött szókincsű nyelvezetet, ontológiákat alkalmaznak.

A genomannotációnak formális definícióját nem találtam az irodalomban, ezért az egyetemi bioinformatika előadások alapján saját magam próbálok felállítani egy olyan fogalmi vázat, amely a bioinformatika területén általánosan alkalmazható.

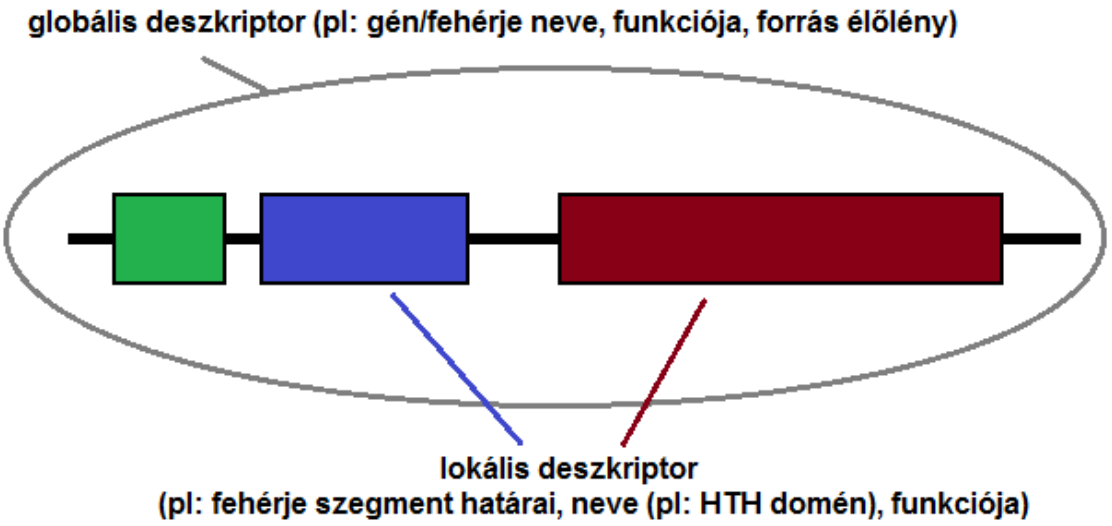
A szerkezet fogalma a bioinformatikában egységek (entitások, szubstruktúrák) és relációk együtteseként fogható fel. Az ilyen szerkezetek leírásai lehetnek egyszerűsítettek (például lehet a szerkezetet csak összetételszerűen, az azt alkotó egységek számával jellemezni), de vannak egyszerű, tulajdonságszerű jellemzések is. Az egységek megengedett neveit, a hozzájuk tartozó tulajdonságok listáját és az egységek között megengedhető relációkat a bioinformatikai ontológiák foglalják össze. Egy egység leírását formálisan megadhatjuk az *entity-attribute-value* adatmodell keretein belül, és ugyanilyen leírások tartoznak a relációkhoz is. Ez az általános megfogalmazás azért fontos, mert kis módosításokkal alkalmazható a bioinformatika fő adattípusaira, vagyis a szekvenciákra (DNS, fehérjék), a 3D szerkezetekre, a hálózatokra és a szövegekre. A genomokat például DNS szekvenciák formájában szokás ábrázolni, itt az egységeket a nukleotidok (A, C, G, T) alkotják, a relációk közül pedig egyedül a láncon belüli szomszédosság relációja szerepel, amit külön nem is tüntetünk fel. A nukleotidokat egy adott ábécé – pl. az IUPAC nomenklatúra – szerint definiált egybetűs kóddal jelenítjük meg, így a szekvencia egy karaktersorozat formájában irható.

Nyers adatokon a genomannotációban egy karaktersorozatot, a genom szekvenciáját értjük. A genom elméleti topológiájaként a számegyenes egész számait, mint pozíciókat képzelhetjük el; a genomszekvenálás (szerkezet-meghatározás) során ezekhez a pozíciókhoz

DOI:10.15774/PPKE.ITK.2014.006

rendeljük a nukleotidok karaktereit. A genom maga állhat egyetlen szekvenciából - ami lehet lineáris vagy cirkuláris - de állhat több ilyenből is. A baktériumoknál általában egyetlen lineáris vagy cirkuláris szekvenciát szoktak megadni, emellett esetleg egy vagy több cirkuláris plazmidszekvencia is szerepel. A genomok szekvenciája lehet komplett (amelynek szekvenciáját megnézték és valamilyen hibaellenőrzésnek is alávetették), de lehet, hogy még csak darabok, úgynevezett *contig*-szekvenciák állnak rendelkezésre, amelyek átfedők is lehetnek. Végül vannak egyedi génszekvenálásból származó rövid DNS adatok is.

Az annotáció folyamata során attribútumokat – a bioinformatikában szokásos elnevezéssel deszkriptorokat – rendelünk a szerkezethez, vagy annak egyes részeihez. Kétféle deszkriptorról beszélhetünk: az egész struktúrára vonatkozó globális deszkriptorokról, illetve annak egy részére vonatkozó lokális deszkriptorokról. (2.3. ábra)



2.3. ábra A deszkriptorok csoportosításának szemléltetése

A deszkriptorok két típusa; a globális deszkriptor az egész struktúrára vonatkozóan ad információt, a lokális deszkriptor csak egy kisebb részére vonatkozóan.

A deszkriptorok forrásai a következők lehetnek:

Emberi tudás. Ezt az annotátorok kötött szókincs, pontosabban ontológiai definíciók segítségével fogalmazzák meg. Formájuk lehet szabad szöveg is.

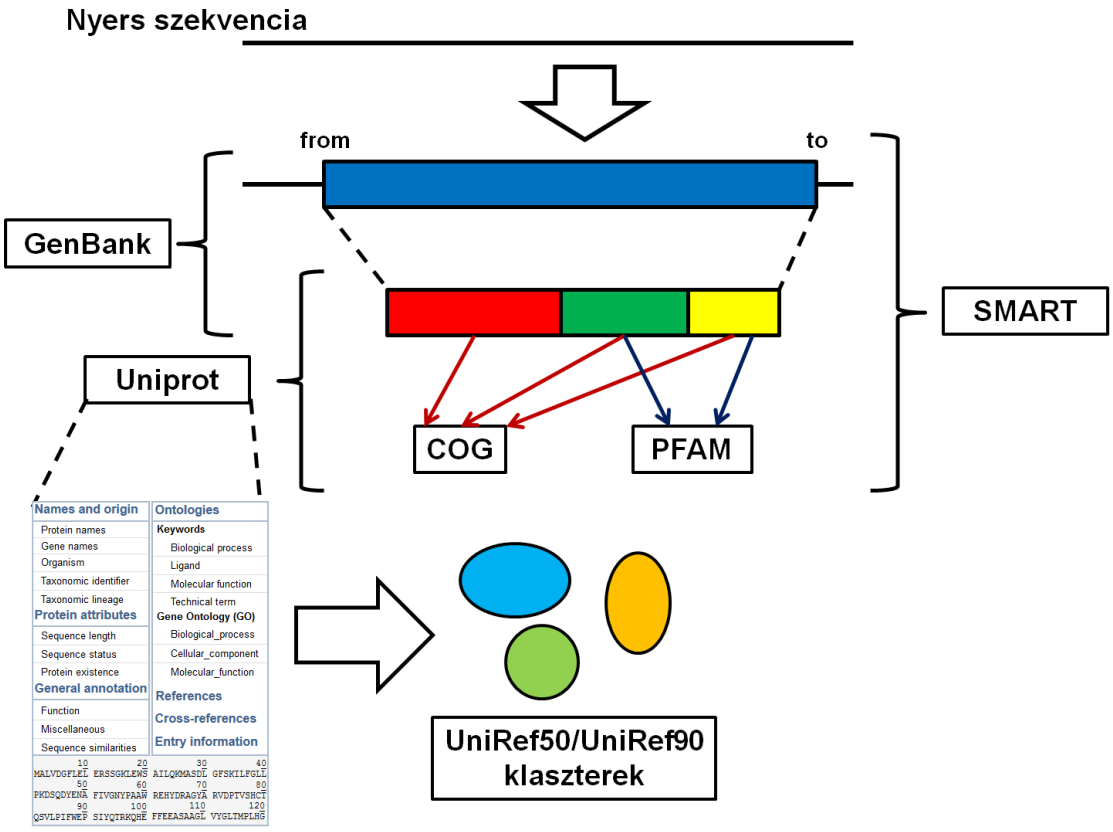
Számítógépes eljárások. Ezeknek két típusa van: vagy valamely más adathoz való hasonlóság alapján, hasonlóságkereséssel definiálunk egy deszkriptort (pl. *putative protease*), vagy a nyers adatokon végzett számítással döntünk el valamit (pl. *low complexity region*).

Adatbázis keresztreferenciák. Ilyenkor a nyers adatot, vagy annak egy részét mutatóval összekötjük egy másik adatbázissal. Ez utóbbi deszkriptorai ugyancsak e három forrásból származnak.

DOI:10.15774/PPKE.ITK.2014.006

A fentiek alapján felvázolhatjuk egy DNS szekvencia annotációjának logikai vázlatát. Egy DNS szekvenciát akkor tekintünk annotáltnak, ha benne megjelöltük a kódoló géneket és egyéb szakaszok, a fehérjét kódoló géneket pedig összekötöttük minél több adatbázissal: az elsődleges fehérje-adatbázisokkal (pl. UNIPROT), a funkció szerint klaszterezett adatbázisokkal, (pl. COG), a szerkezet alapján klaszterezett adatbázisokkal (pl. PFAM)… stb. Ez a vázlat két közvetlen következtetést sugall:

- Általában az annotáció nem „teljes”, sok új fehérjének még nincs rekordja az egyes adatbázisokban, vagy ha van is, azok annotálása nem teljes.
- Az annotációk gyorsan változnak, hiszen az egyes háttér-adatbázisokat rendszeresen frissítik. Ezért egy szekvencia vagy genom annotációja elvben sem lehet teljes, mert a keresztreferenciákkal hozzákapcsolt adatbázisok révén az annotáció tartalma is állandóan aktualizálódik.



2.4. ábra A genomannotáció alapvető lépései.

A fenti ábrán egy genomannotáció lépéseit láthatjuk a bioinformatikai adatbázisokra vetítve. A nyers szekvencián először megállapítjuk a gén helyét. Ezután megkeressük a gén szubstruktúráit (domének), és hozzátársítjuk az eddig ismert doménccsaládokhoz (COG, PFAM). Az így megismert fehérjét beillesztjük a UniProt adatbázisba. Ezután megvizsgáljuk azon UniRef klasztert, amelyik tartalmazza az új fehérjét.

DOI:10.15774/PPKE.ITK.2014.006

A fenti idealizált állapotot csak egy jól szervezett és frissíthető integrált adatbázis képes megközelíteni, a jelenleg nyilvánosan hozzáférhető adatbázisok azonban nem ilyenek. A gyakorlatban annotált genomon olyan genom szekvenciákat értünk, amelyek egy adatbázisban vannak elraktározva. Az NCBI (National Center for Biotechnology Information) tartja fent a legismertebb ilyen gyűjteményt. Ebben az adatbázisban a bennünket érdeklő bakteriális genomokra a következő kategóriák léteznek:

Complete genomes; Teljes vagy annotált genom alatt olyan genomszekvenciát értünk, melyekben az összes gén helye meg van határozva és egy részük - általában a többségük - funkciós leírással is rendelkezik. Ennek szekvenciája validálva van.

Draft genomes; Ezek több, össze nem állított szekvencia-darabból, úgy nevezett *contigok*ból állnak. Némelyek annotáltak, tehát van *ptt* file is hozzájuk (lásd lejjebb). A *draft* genomra nincs elfogadott magyar szó, leginkább a „félkész genom" vagy a „feldolgozás alatt álló genom" kifejezések írják le a fogalmat.

Ezeken felül természetesen léteznek a régi típusú DNS szekvencia rekordok is, ezeket a GenBank tartalmazza. A GenBank annotációs része tartalmazza a fehérjeszakaszok szekvenciáját, a gének helyeit… stb.

A fentiek alapján megfogalmazhatjuk az annotáció fogalmát az adatok formátuma szempontjából is. A nyers adatokat a legegyszerűbb szekvenciaformátumokban szokás letétbe helyezni, mint pl. a FASTA illetve a konkatenált FASTA formátum: ennek egyetlen annotációs sorában az adat származásáról, a kísérlet számáról és/vagy a DNS származásáról (a szervezet biológiai nevéről) találunk információt, gyakran csak *ad hoc* megfogalmazás formájában. Amennyiben egy ilyen DNS rekord bekerül egy nyilvános adatbázisba, pl. az NCBI-hoz, akkor egy megváltoztathatatlan azonosítót kap, ami megjelenik a FASTA annotációs sorában.

A kész genomokat a következő adatformátumokban szokás megadni:

faa: Konkatenált FASTA file, ami egy bakteriális kromoszóma vagy plazmid génjeinek aminosav-szekvenciáit tartalmazza, minden szekvenciához egy annotációs fejléccel, ami a gén fontosabb azonosítóit, a valószínűsíthető funkcióit és a forrás baktérium nevét tartalmazza. (II. Melléklet)

ffn: Az előzőhöz hasonló konkatenált FASTA file, csak a gének nukleinsav-szekvenciáit tartalmazza a génre vonatkozó információk nélkül. (III. Melléklet)

fna: A baktérium DNS-ének teljes nukleinsav-szekvenciája mindenféle csoportosítás és megszakítás nélkül, FASTA formátumban. A fejléc a baktérium nevét és a szekvenálás állapotát tartalmazza.

DOI:10.15774/PPKE.ITK.2014.006

gbk: Az adott bakteriális kromoszóma vagy plazmid GenBank rekordja, ami minden információt tartalmaz az adott bakteriális genommal kapcsolatban: baktérium neve, teljes taxonómiája, azonosítói, folyóirat referenciák, a szekvencia teljes elemzése génekre és egyéb jellegzetességekre való tekintettel, és maga a teljes szekvencia. Ez a formátum áttekinthetőbb az ember számára, de program általi feldolgozása nehezebb a többi említett fájltípushoz képest. (IV. Melléklet)

ptt: Táblázatos felépítésű adatfájl, ami az adott bakteriális kromoszóma vagy plazmid génjeit tartalmazza a DNS-en való elhelyezkedés sorrendjében. Soronként egy gén és a hozzá tartotózó adatok (például: pontos elhelyezkedés a szekvenciában, azonosítók, elnevezései, valószínűsíthető funkció… stb.) vannak. Információ kinyerés szempontjából ideális, mert könnyen feldolgozható és kevés számunkra felesleges adatot tartalmaz. (V. Melléklet)

A félkész genomok a nyers adatok és a kész genomok közötti készültségi fokban vannak, tehát a *contig*okról néha találunk annotációt, néha nem.

A bakteriális genomok átlagosan mintegy 5 millió bázispár hosszúak, és bennük mintegy 3-5000 gén van. Az NCBI adatbázis jelenlegi állapota szerint egy jólannotált bakteriális genomban is mintegy 25% nem annotált gén van, ezeket „*gene of unknown function*”, „*hypothetical protein*” stb. címkékkel látják el. Általában a baktériumok alapgénjeit annotálják részletesen, ezek a gének a legtöbb baktériumban megvannak, és alapfunkciókat látnak el. A „járulékos géneket” (*shell*), amelyek a baktériumok speciális, gyakran csak egyetlen fajban megtalálható funkcióit látják el, már sokkal kevésbé annotálják. Sokszor pedig vannak igen jó becslések ezek funkciójára, viszont az annotátorok illetve a nyilvános adatbázisok fenntartói kerülni akarják az annotációs hibákat; ennek következményeként sok, egyébként annotálható gént inkább a hipotetikus kategóriában hagynak.

A genomannotáció munkafolyamataira két megközelítés létezik: egy kiválasztott genom teljes annotációja és egyes funkcionális egységek, alrendszerek (pl. metabolikus útvonalak) annotációja több genomban.

Az első esetben kiválasztunk egy genomot, amelynek minél több, eddig ismeretlen génjének funkcióját próbáljuk felderíteni biológiai vizsgálatok vagy szekvencia adatbázis keresések segítségével. Ennek előnye, hogy csupán egy faj kellő ismerete elég lehet gének megismerésére, viszont rá vagyunk utalva az eddigi génadatbázisok adataira, amely hibás annotálásra ad lehetőséget. Például egy hasonlóságalapú keresés esetén könnyen kaphatunk fals pozitív eredményt, mivel ez a módszer csak azt mutatja meg, hogy az eddig ismert gének közül melyikhez hasonlít leginkább. Ez nem garantálja, hogy ténylegesen olyan funkciójú a gén, mert elképzelhető, hogy egy eddig nem dokumentált feladatot lát el a vizsgált génszakasz.

DOI:10.15774/PPKE.ITK.2014.006

A természet diverzitásának következtében a keresések során ugyancsak könnyen kaphatunk fals negatív eredményt, ha egy eddig ismeretlen alléllal találkozunk.

Az alrendszerek annotációja esetén nem egy genomot választunk ki, hanem egy meghatározott alrendszert, ami egy funkcionális szabálygyűjtemény, amely meghatároz egy biológiai folyamatot vagy struktúrát.[23] Miután megismertük az alrendszerrel kapcsolatban álló géneket, ezeket a géneket keressük az eddigi ismert genomok szekvenciáiban. A talált új gének funkciójának validálásában segítségünkre van az alrendszer szabályrendszere, amely segít kiszűrni mind a fals pozitív, mind a fals negatív eredményeket. Viszont ez a módszer sem ad biztos eredményt, mert előfordulhat eddig nem ismert variációja az alrendszernek, illetve kis génszámú alrendszer esetén a más génekkel való véletlen hasonlóság is hibát okozhat.

2.3. A genomannotáció eszközei

2.3.1. Többszörös illesztés és ClustalW

A DNS szekvenciákban lévő hasonlóságok felderítésének egyik népszerű módja a vizsgált szekvenciák illesztése. A páros illesztés lényege, hogy a két szekvenciánkat hézagok beiktatásával úgy rendezzük, hogy egy előre definiált távolságfüggvényt minimalizáljanak. Ez az illesztés lehet globális illetve lokális illesztés. Globális esetben a szekvenciák egész hosszát egyben vizsgálva próbáljuk őket minél jobban hasonlóvá tenni, míg lokális illesztés esetén előnyben részesítjük a hosszabb hasonló szakaszokat és az összefüggő hézagokat. Ez utóbbi azért hasznosabb, mert felismerheti a mindkét szekvenciában meglévő struktúrákat és doméneket. A két legismertebb algoritmus a globálisan illesztő *Needleman-Wunsch* [24] és a lokális *Smith-Waterman* algoritmus [25].

Az illesztés közben használt távolságfüggvény egy *helyettesítési* (*substitution*) mátrixon alapszik, amely minden aminosav párra tartalmazza az átváltási költséget. Mivel 20 aminosav létezik, ezért ez a mátrix egy 20x20-as, szimmetrikus mátrix. A mátrixnak két főbb típusa létezik: az evolúciós kapcsolatokon alapuló PAM (*Percent Accepted Mutation*) és a PROSITE adatbázis szekvencia hasonlóságain alapuló BLOSUM [26] (*Blocks Substitution Matrix*) mátrix. Mindkét mátrixnak több altípusa van, amelyek funkcióit a 2.1. táblázat mutatja be.

2.1. táblázat A helyettesítési mátrixok típusai

PAM		BLOSUM
A számok azt jelzik, hogy a PAM1 mátrixot hányszor szorozták össze magával.		A számok az jelzik, hogy hány százalékosan hasonlító szekvenciák segítségével készült.
PAM 40	rövid, nagyon hasonló szekvenciákhoz	BLOSUM 90
PAM 120	általános illesztéshez	BLOSUM 62
PAM 250	távoli hasonlóságok kimutatására	BLOSUM 30

DOI:10.15774/PPKE.ITK.2014.006

Mivel számunkra nem kettő, hanem több tucat szekvencia összehasonlítására van szükség, ezért többszörös illesztést kell alkalmazni. A többszörös illesztés egy kétdimenziós táblázat, melynek sorai a szekvenciákat tartalmazza, oszlopai pedig a pozíciókat jelképezi. (lásd 2.5. ábra) A többszörös illesztés elkészítésének több módja van, de mind megegyezik abban, hogy az összes lehetséges páros illesztésből kiindulva kísérel meg egy közös illesztést kialakítani. A legtöbb esetben progresszív módszereket használnak, melyek valamilyen heurisztika segítségével próbálnak „relatíve rövid" idő alatt jó rendezést (bár nem biztos, hogy optimálisat) találni. Ezért figyelnünk kell arra, hogy a feladatunknak megfelelő heurisztikát válaszunk az illesztés elkészítéséhez.

A *ClustalW* [27] napjaink egyik legelterjedtebb és leginkább használt többszörös illesztést készítő algoritmusa. Bár minden többszörös illesztés alapja a páronkénti illesztés, azonban lényeges különbség közöttük, hogy ezeket milyen módon próbálják egyesíteni egy illesztésé, mert a használt metódustól függően különböző eredményt kapunk és ezen eredmények különböző biológiai helyességgel bírnak. Például a mohó algoritmusok nagyon hasonló szekvenciák esetén viszonylag jó, biológiailag is informatív illesztést eredményeznek, kevésbé hasonló szekvenciák esetén viszont gyakran használhatatlan, biológiai szempontból nézve semmitmondó illesztést eredményeznek. A *ClustalW* előnye, hogy progresszív módon, megfelelő súlyozást alkalmazva készíti el a többszörös illesztést. Erre utal a nevében a W, ami az angol *weighted* (=súlyozott) szóból ered.

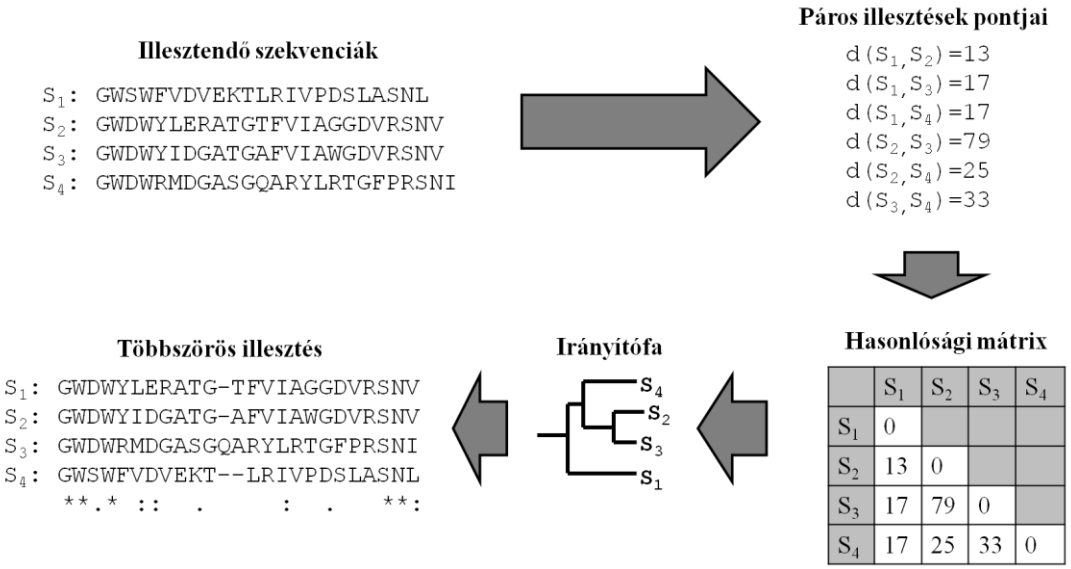
```
RQTRRLLLDLKPDGHAMPAYFDTCSHDGYVRVSATQLASITLALLYAACDERVLLGLPACHAGHCEWIDTNWR--PPMTFGWAWLT
-------MDLKPDGHAVPAYFDTCSHDGYVRVSATQLASITLALLYAARDERVLLGLPACHAGHCEWIDTNWR--PPMTFGWAWLT
GRRSHVAQRLAP-ASSAPA---AAGRDGYVRLSLPQLARVSLAPICAAADDQILLGIGALRAGYCEWIDAHGA--APATLGWSWFV
-----MAQRLAP-ASSAPA---AAGRDGYVRLSLPQLARVSLAPICAAADDQILLGIGALRAGYCEWIDAHGA--APTTLGWSWFV
-----MAQRSAP-ASSAPA---AAGQDGYVRLSLPQLARVSLAPTFAAADDQILLGICALRAGYCEWSDAHWP--APTSLGWSWFV
--------MTSPLLHPVPG----PSPDGYVRLSEGALAALALDHVASGLDPSLLNAIDARLAGYTEWHRPAGAGVAYVTVGWDWYL
--------MTSPLLHPVPG----PSPDGYVRLSEGALAALALDHVASGLDPSLLNAIDARLAGYTEWHRPAGAGVAYVTVGWDWYL
--------MTSSLLHAVRG----PSPDGYVRLSESALAALALDHVASGLDPALLSAIDARLAGYTEWQRPAGPGVAYVTVGWDWYL
--------MNSPWLHRFRG----SSTDGYVRLPMHAFAELRLVHVSSGIDSGLLSDIDARIAGYTEWERPSSAGAAHLTVGWDWYI
--------MNSPWLHRFRG----SSTDGYVRLPMHAFAELRLVHVSSGIDSGLLSDIDARIAGYTEWERPSSAGAAHLTVGWDWYI
--------MTPPLLHPYCS----PSADGYVRLPLHAFAGLELVHIASGLDPGLAEGLVPDLAGFTEWQRPASPGYAHLTVGWDWYL
-----------------MR----LSPDGYVRLTLEQFQWIPLVHLLSGLDHDEYGASQTHISGYTEWVSETAP---VITLGWDWRM
-MTVQNGIAVKPQGYGFMR----LSPDGYVRLTLEQFQWIPLVHLLSGLDHDEYGASQTHISGYTEWVSETAP---VITLGWDWRM
--------------------------------MNENQLMSYPLQHIISTVESRHT----IFYYGFTEWATSQTP---ALSTGWDWEL
--------------------------------MNENQLMSYPLQHIISIVESRHT----IFYYGFTEWATSQTP---ALSTGWDWEL
----------MVSIPPHFS----ISTDGFIRMNENQLMSYPLQHIISTVESRHT----IFYYGFTEWATSQTP---ALSTGWDWEL
----------MVTIPPDFP----ISTDGFIRMNEIQLMNYPLQHLISIVETTQI----ILYCGFTEWATSRSP---ALSTGWDWEF
-----MHNSRPEYLKEVLN----VSHGGYVRMSRSAFMELPLSHFISGLDEDPALPDESTISGYTEWLSVCVP---IITVGWDWRL
                                :    :    *    :  :            *. **           : ** *
```

2.5. ábra Egy többszörös illesztés részlete

A legalsó sorban szereplő karakterek az adott sor konzerváltságának mértékét írják le. A képen egy teljesen sima szövegszerkesztőben megjelenített, szöveges fájlban tárolt illesztés látható. Sokkal látványosabb eredmény érhető el specifikus megjelenítő programok (például: Jalview [28]) segítségével. (I. melléklet)

DOI:10.15774/PPKE.ITK.2014.006

A *ClustalW* algoritmus a legtöbb többszörös illesztést készítő algoritmushoz hasonlóan először megkonstruálja az összes lehetséges páronkénti illesztést. Ezen illesztésekhez egy távolság értéket rendel, attól függően, hogy a két szekvencia mekkora mértékben hasonlít egymásra. Ezeket az értékeket egy hasonlósági mátrixba gyűjti. A mátrix alapján egy irányítófát készít *neighbor-joining* metódus használatával. A *ClustalW* algoritmus a többszörös illesztést az irányítófa elágazásainak sorrendjében végzi, először a két leginkább hasonló szekvenciát választja ki, és illeszti őket. Ezután az egyre távolabbi szekvenciákat illeszti az eddigi illesztésünkhöz, szükség esetén megfelelő számú hézaggal kiegészítve. Ezen algoritmus segítségével viszonylag különböző szekvenciák esetén is helyes többszörös illesztés kapunk.



	S1	S2	S3	S4
S1	0			
S2	13	0		
S3	17	79	0	
S4	17	25	33	0

2.6. ábra A ClustalW algoritmus alaplépései

A példában *rsaM* gének fehérjeszekvenciáinak egy részlete található. Mivel ezek gének körülbelül 100 bázispár hosszúak, ezért az áttekinthetőség kedvéért csak egy rövidebb szakasszal illusztráltam az algoritmus működését.

2.3.2. Rejtett Markov Modell

A *Rejtett Markov Modell* (HMM = *Hidden Markov Model*) alapú keresések az egyik legkeresettebbek a bioinformatikai annotálási feladatokban. Magas megbízhatóságuk, és számítógépen történő könnyű implementációjuk teszi őket rutin eszközökké. [29-31]

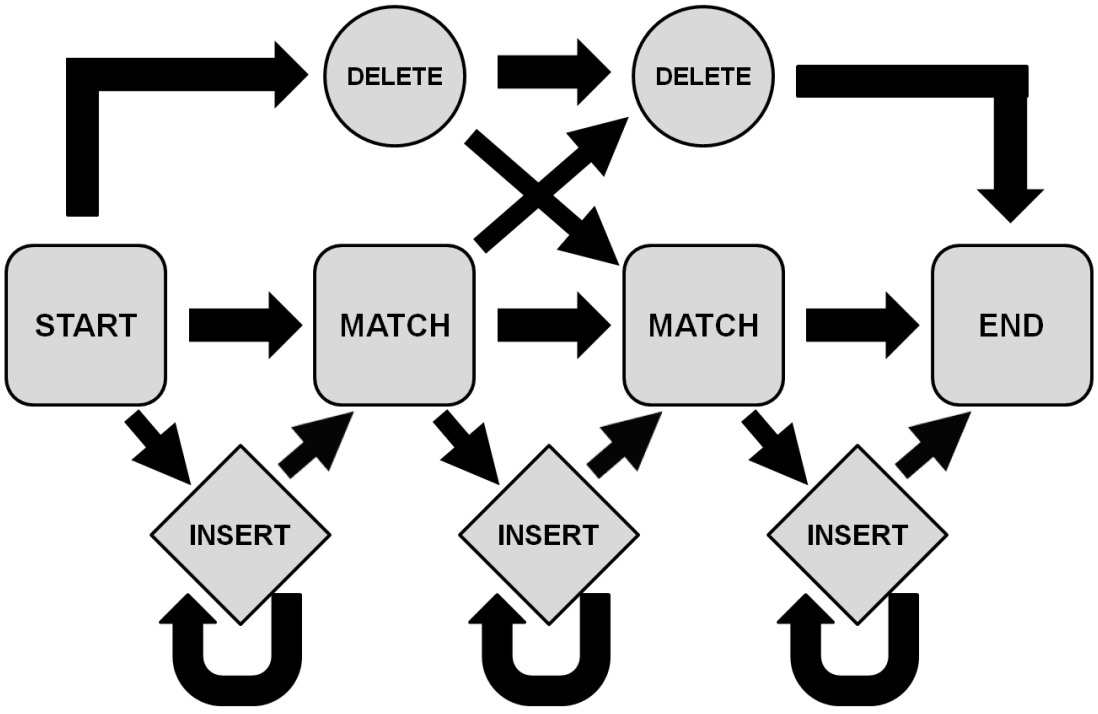
A modell alapja egy automataszerű struktúra, így néhány adat elég a definiálásához:

- *Állapotok halmaza*: $\boldsymbol{Q} = \{q_0, q_1, q_2 \dots q_n\}$
- *A kimeneti ábécé*: $\boldsymbol{\Sigma} = \{v_1, v_2, v_3 \dots v_m\}$
- *Kezdő állapot*: $\boldsymbol{\pi}(i)$ annak a valószínűsége, hogy a kezdő állapot q_i

DOI:10.15774/PPKE.ITK.2014.006

- *Állapot átmenet:* $\boldsymbol{A} = \{a_{ij}\}$ ahol a_{ij} annak a valószínűsége, hogy ha most a q_i állapotban vagyunk, akkor a következő lépésben a q_j állapotban leszünk.
- *Kilépési valószínűség*: $\boldsymbol{B} = \{b_j(k)\}$ ahol $b_j(k)$ annak a valószínűsége, hogy ha most a q_i állapotban vagyunk, akkor egy v_k jelet generálunk

A *rejtett Markov modell* készítésének alapja a forrás szekvenciák többszörös illesztése, melynek oszlopait egyesével véve folyamatosan felépítjük a modell alapjául szolgáló automatát. Az illesztés nyomán minden oszlophoz készül egy *egyezőségi* (**M** = *match state*) *állapot*, ami 20 valószínűséget tartalmaz (aminosavanként külön-külön), az illesztés oszlopaiban szereplő aminosavak előfordulási száma szerint. Továbbá az illesztés szerinti törlések és beszúrások *törlés állapotként* (**D** = *delete state*) és *beszúrás állapotként* (**I** = *insertion state*) kerülnek a modellbe. Minden **M** állapothoz tartozik egy **D** állapot az adott aminosav hiányzásának szimulálására. Ezenfelül pedig az **M** állapotok között egy önrekurzív **I** állapot található, amely garantálja a megfelelő méretű hézag beillesztésének lehetőségét. Az előbb említett állapotok irányított élekkel köthetők össze az állapot változási valószínűséggel súlyozva. (2.7. ábra) Minden állapot rendelkezik egy visszatérési értékkel, amely értékek az automatában való lépések folyamán a kimenetet eredményezik. A modell „rejtettsége" abban nyilvánul meg, hogy a belső állapotok sorrendjét nem ismerjük, csak az állapotok által generált kimenetet (ami maga a szekvencia lesz).



2.7. ábra A HMM állapot átmeneti diagramjának egyszerűsített ábrája

A különböző típusú állapotok különböző módon vannak jelölve; az egyezések téglalappal, a beszúrások rombusszal, a törlések körrel.

DOI:10.15774/PPKE.ITK.2014.006

Ha elkészült a HMM profilunk, akkor már csak egy keresést kell elvégezni az adatbázisunkon a profil alapján. Ez a művelet egy listát készít a profilhoz legjobban hasonlító szekvenciákból, jelezve a hasonlóság értékét. Egy példa a pontozási módra:

$$S = log_2 \frac{P(szekvencia|HMM)}{P(szekvencia|null)}$$

Ahol *P(szekvencia|HMM)* annak a valószínűsége, hogy a szekvencia a profilhoz tartozik és a *P(szekvencia|null)* annak a valószínűsége, hogy a szekvencia az úgynevezett null modellhez tartozik. A null modell lényegében olyan véletlen szekvencia, aminek komponensei egy megfelelő hosszú független egyenletes eloszlásnak felelnek meg. Ez a két valószínűség a modellbe beépített egyes állapotokhoz tartozó valószínűségek alapján számíthatóak ki. A HMM alapú keresést végrehajtó szoftverek egy másik hasonlósági értéket is megadnak: az *e-value*-t. Ezt a számot a program a szekvencia hasonlósági értékéből (*score*) számolja ki, ez viszont függ az adatbázis méretétől is, ugyanis az *e-value* azt mutatja meg, hogy egy adott *score* esetén mennyi a fals-pozitív eredmények várható értéke a *score* fölött (minél kisebb az *e-value* annál biztosabb a találat).

Ennek a keresési módszernek komoly előnyei vannak. Először is nagyon fontos, hogy akár egy teljes fehérjecsaládot reprezentálhatunk egy HMM profillal. Ez a keresések megbízhatóságára nyilván komoly hatással van, hiszen sok annotációs eljárással ellenétben nem egy szekvenciával szemben vizsgáljuk a benyújtott szekvenciánkat. További fontos előnye a HMM alapú modellezéseknek, hogy több profilt is egybefoghatunk, létrehozva ezzel egy HMM könyvtárat. Végül van egy hátránya is a HMM profiloknak: a felhasználónak lehetősége van a túltanításukra. Ilyen esetekben az elkészült profil lényegében nem egy családot képvisel, tehát érzékeny keresési eljárás megvalósításába nem vonható be.

2.3.3. BLAST

A szekvencia-összehasonlító algoritmusok egyik legnagyobb úttörője a BLAST (*Basic Local Alignment Search Tool*), [32] melynek alapja a heurisztikus szekvencia-összehasonlítás. A korábban használt, kimerítő keresést alkalmazó illesztő algoritmusok (mint például a Smith-Waterman) megtalálták ugyan mindig az optimális megoldást, de hosszabb szekvenciák használata esetén a keresési idő drasztikusan nőt, ami a hatalmas szekvencia adatbázisok vizsgálatát gátolta. A lefutási idő problémáját a BLAST algoritmus heurisztika alkalmazásával oldotta meg.

Magának az algoritmusnak a bemenete egy szekvencia (ezt hívjuk célszekvenciának). A BLAST algoritmus minden adatbázisban szereplő szekvenciához egy hasonlósági értéket rendel a célszekvenciához illeszkedés alapján.

DOI:10.15774/PPKE.ITK.2014.006

A BLAST algoritmus lépései

1) **Az alacsony komplexitású régiók és szekvencia ismétlődések eltávolítása**

   Az alacsony komplexitású régió azt jelenti, hogy a célszekvenciánk hosszú, monoton részeket tartalmaz, ami később magas pontszámot okozva megzavarja az algoritmust. Ezeket a szakaszokat a programok általában külön megjelölik **X** (aminosav) vagy **N** (nukleinsav) betűvel a szekvencián.

2) k betűs szólista készítése

A célszekvenciából képezik az összes lehetséges k hosszú részszekvenciát. Aminosav szekvenciák esetén általában 3, míg nukleinsav esetén 11 hosszú szavakat használnak. (2.8. ábra)

3) Lehetséges egyező szavak listázása

A BLAST algoritmus minden előző lépésben készített szólista minden tagjához egy számot rendel, és csak a magas értékkel rendelkezőkkel foglalkozik a későbbiekben. A szám előállításához a szót betűnként összehasonlítjuk minden lehetséges k hosszú szóval. Minden összehasonlítás egy értéket eredményez valamilyen pontozó mátrix alapján, és ezeket összeadva kapjuk az egész szó pontszámát. Ha pontszám magasabb a megadott határértéknél, akkor megtartja, ha nem elveti.

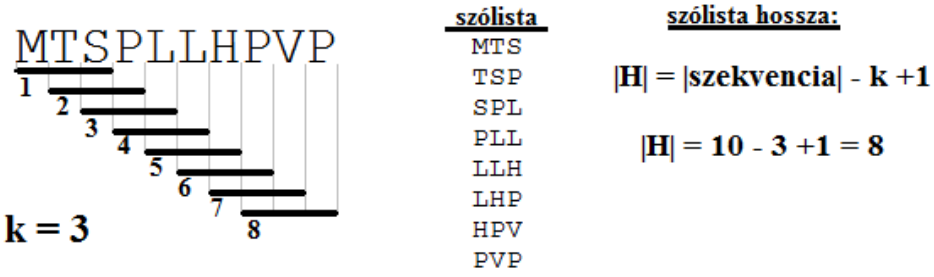
4) A magas pontszámú szavak keresőfába rendezése

Ez a lépés lehetővé teszi a program számára, hogy majd az adatbázis szekvenciáival gyorsan össze tudja hasonlítani őket. Erre a célra egy hash-tábla használható, mert gyors hozzáférési idővel rendelkezik.

5) A 3) és 4) ismétlése a célszekvenciából készült szólista minden elemére

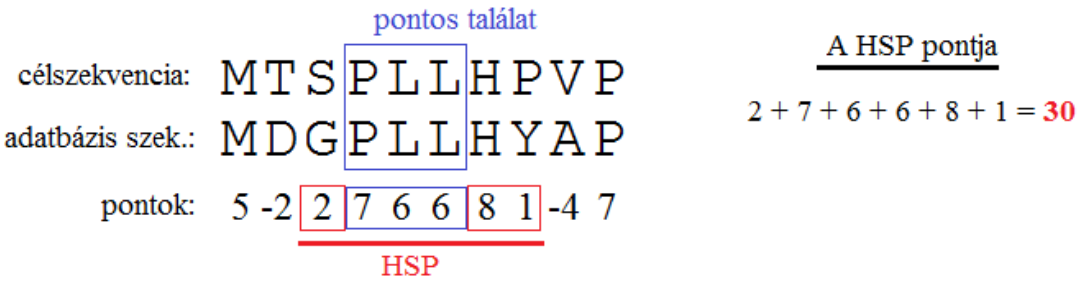
6) Az adatbázis szekvenciákban pontos egyezést keresünk a szűkített szólistánkkal

A szólista szavait minden lehetséges helyen hozzápróbáljuk az adatbázis szekvenciákhoz, és ha pontos egyezést tapasztalunk, akkor ezt a részletet *seed*ként használjuk egy lehetséges hézagnélküli illesztéshez.



2.8. ábra A BLAST algoritmus szólistája k=3 esetén

DOI:10.15774/PPKE.ITK.2014.006



2.9. ábra A BLAST algoritmus találat kiterjesztése

7) A pontos egyezések kiterjesztése HSP-vé

A pontos egyezéseket ezután megpróbáljuk kiterjesztéssel megnövelni. Ez azt jelenti, hogy mind jobb, mind bal irányba elindulva egyesével hozzávesszük a szomszédos elemeket, egészen addig, míg az így kapott hosszabb szakasz pontszáma nagyobb, mint az új elem nélküli pontszáma. Ezeket a kiterjesztett régiókat HSP-nek (*High-scoring Segment Pair*) hívjuk. (2.9. ábra)

8) Magas ponttal rendelkező HSP-k listázása

Összegyűjtjük az összes olyan keletkezett HSP-t, aminek pontszáma nagyobb, mint egy tapasztalati úton meghatározott vágási határérték. Ennek az értéknek a meghatározása véletlen szekvenciák összehasonlításán alapszik.

9) A HSP-k szignifikanciájának kiszámítása

A szignifikancia kiszámításához a *Gumbel extrém érték eloszlás*t (*Gumbel Extreme Value Distribution*) használjuk. Annak a valószínűsége, hogy az **S** megfigyelt pontszám nagyobb egy **x** értéknél:

$$P(S \geq x) = 1 - \exp\left(-e^{-\lambda(x-\mu)}\right)$$

Ahol $\mu = \frac{log_2(K*m'*n')}{\lambda}$ $\mu = \frac{log_2(K*m'*n')}{\lambda}$, m' és n' pedig az effektív hosszúsága a célszekvenciának és az adatbázis szekvenciáknak. A K és λ paramétereket a célszekvencia és az adatbázis szekvenciák illesztése révén kapjuk.

10) HSP régiók összefűzése illesztésé

Néha előfordul, hogy kettő vagy több HSP régió egy adatbázis szekvenciában összefűzhető egy hosszabb illesztésé. Ez további bizonyítékkal szolgál a célszekvencia és az adatbázis szekvencia kapcsolatára.

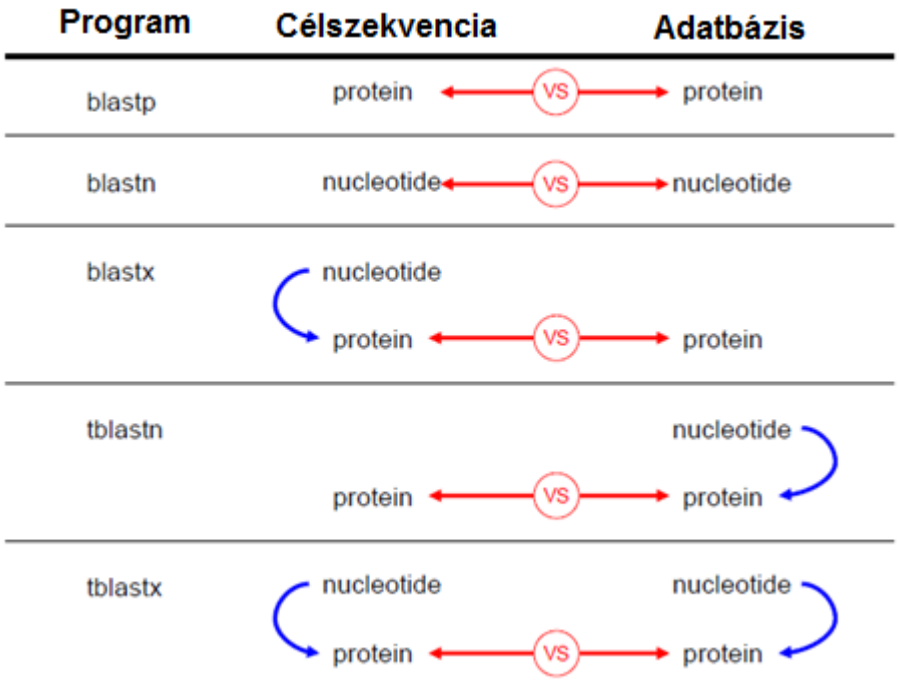
11) A célszekvencia és a szignifikáns adatbázis szekvenciák lokális illesztése

A lokális illesztés elvégzéséhez a Smith-Waterman algoritmust használjuk minden szignifikáns találattal rendelkező adatbázis szekvencia esetén.

12) Minden megadott határértéknél jobb találat listázása

DOI:10.15774/PPKE.ITK.2014.006

A BLAST algoritmusnak több altípusa is van, attól függően, hogy milyen típusú a célszekvencia, és milyen típusú szekvenciák vannak az adatbázisban. (2.10. ábra)



2.10. ábra A BLAST algoritmus típusai

2.4. A genomannotáció típusai

2.4.1. A szerkezet alapú genomannotáció

Ez alatt a DNS-szekvencia egyszerű szerkezeti elemeinek (szakaszainak) azonosítását értjük, amelyet a szekvencia alapján végzünk el. A fehérje kódoló gének meghatározása baktériumokban nem bonyolult, de nem triviális probléma. Többféle módszer is létezik a genomannotációra, melyeknek egyik csoportja a szerkezet alapú. Ez a génannotáció kizárólag a szekvencia karakterisztikáját használja ki és mintafelismerésen alapul. Ezek a minták sokfélék lehetnek, a teljesen pontos megegyezéstől, a bonyolult reguláris kifejezésekkel leírt mintákig. A pontos megegyezésre jó példa az ORF (*Open Reading Frame*) keresés, amelynek alapja a start (ATG) és stop kodonok (TAA, TGA, TAG) keresése a DNS szekvenciában, majd a talált tripletek egymás mellé illesztése. Mindehhez elvben csak a kodontáblázatot kell ismerni, de a kapott szekvenciák közül ki is kell választanunk azokat, amelyek rendelkeznek a gének, például a bakteriális gének karakterisztikáival. Baktériumoknál ezt a GLIMMER programmal szokták elvégezni, amelyik egy rejtett Markov-lánc típusú program.

DOI:10.15774/PPKE.ITK.2014.006

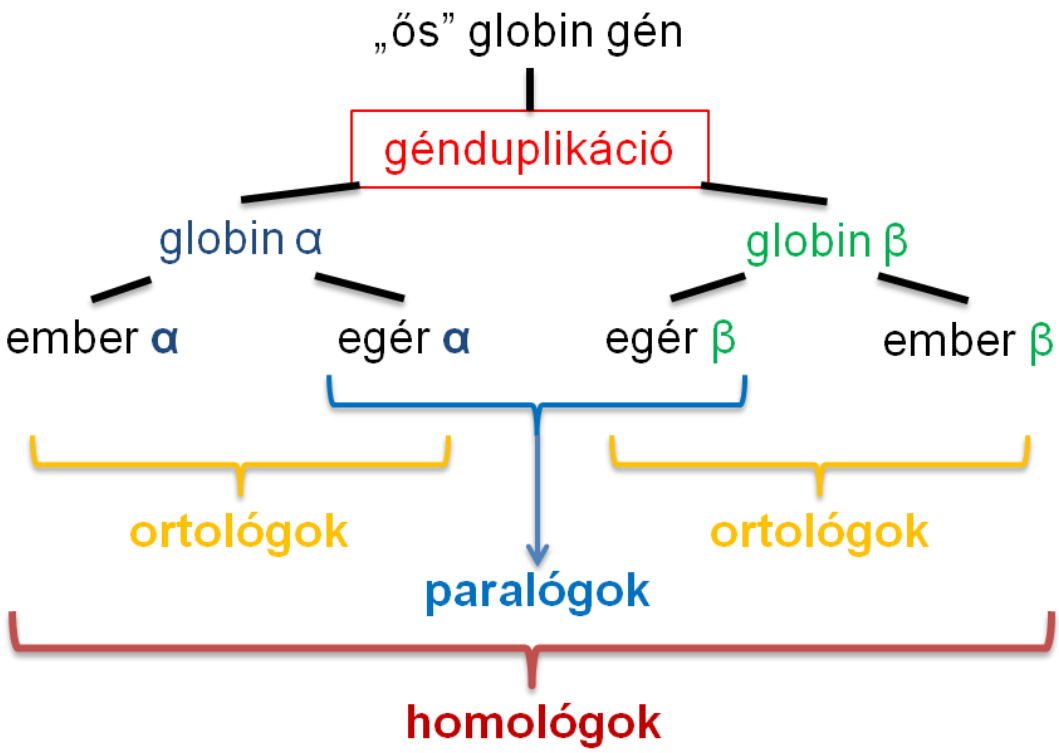
2.4.2. A funkcionális genomannotáció

Miután sikeresen azonosítottuk a genomban a gének és más szekvencia elemek helyét, az annotáció következő lépéseként meg kell határoznunk a molekuláris funkciót és a biológiai szerepet. Elsősorban a gének és az általuk termelt fehérjék azonosításán van a hangsúly. A funkció feltárásához szükséges információt a gének már létező genomikus adathalmazokkal való kapcsolatai alapján ismerjük meg. A kapcsolat lehet hasonlóság ismert funkciójú génnel, lehet közös genomikus szomszédság vagy szabályozójel. A legszigorúbb funkcionális génannotáció a kísérletezés útján történő vizsgálat. Ennek az az előnye, hogy eddig ismeretlen funkciók is felismerhetők vele, és a prediktált szerep valószínűsége a legtöbb esetben magas. Hátránya viszont az, hogy sokkal időigényesebb, mint hogyha a már ismert nyilvános funkció adatbázisokat használnánk. Bár ezeknek az adatbázisoknak a mérete és információ tartalma rohamosan nő, a genomokban szereplő összes gén funkciójának még csak kis részét tudjuk velük lefedni. Ha ehhez hozzá vesszük, hogy a nem kódoló DNS szakaszokról még alig van információnk, egy élőlény teljes genomjának annotációja még távoli célnak tűnik.

2.4.3. A homológ alapú funkcióbecslés

A funkció meghatározásának egyik klasszikus alapja a gének közötti evolúciós kapcsolat. Ezek a módszerek az úgynevezett homológián alapulnak: a keresett gént összehasonlítjuk az adatbázis már ismert működésű génjeivel, és ha szignifikáns hasonlóságot találunk, akkor feltételezhetjük, hogy az ismeretlen génnek is azonos a szerepe. Ez a génhasonlóság több fajt is érinthet, de a kísérleti tapasztalat azt mutatja, hogy gyakran teljesen különböző élőlények esetén is az azonos gének szerepe megegyezik. Ennek a módszernek azonban több nehézsége is van. A génszekvenciánkról nem tudhatjuk, hogy szerepel-e a funkciója az adatbázisban, vagy egy eddig nem ismert szerepkörrel rendelkezik. Ez sok esetben megnehezíti annak az eldöntését, hogy a módszerünk által meghatározott hasonlóság valóban tekinthető-e szignifikánsnak, vagy csak véletlen egyezést tapasztaltunk. Alapvető probléma azonban, hogy a homológ gének szerepe sem biztos, hogy teljes mértékben megegyezik, mert a gén viselkedésére hatással lehetnek a környező gének és a bekövetkezett mutációk is. Ebből a szempontból a homológokat két csoportra oszthatjuk: *ortológ* és *paralóg*. Két gént ortológnak nevezünk, ha két különböző fajban találhatóak, és egy közös ősgénből származnak, mely a két faj közös ősében volt jelen. Ezen gének ugyanazt a funkciót szolgálják a két fajban. Két gént paralógnak nevezünk, ha ugyanabban az organizmusban találhatóak, és egy közös ősgénből génduplikáció és azt követő divergens evolúció útján alakultak ki. Többnyire különböző, de egymással összefüggésben lévő funkciójuk van. (2.11. ábra) [33]

DOI:10.15774/PPKE.ITK.2014.006



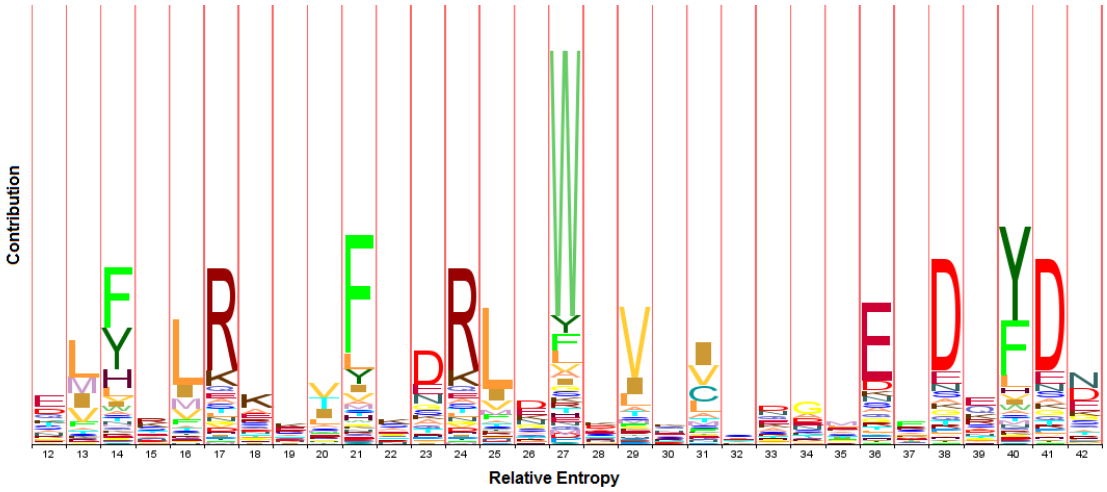
2.11. ábra A szekvenciák homológ kapcsolatai

Két gént ortológnak nevezünk, ha két különböző fajban találhatóak, és egy közös ősgénből származnak, mely a két faj közös ősében volt jelen. Két gént paralógnak nevezünk, ha ugyanabban az organizmusban találhatóak, és egy közös ősgénből génduplikáció és azt követő divergens evolúció útján alakultak ki.

2.4.4. A fehérje domének

A fehérjék összetett háromdimenziós struktúrák, melyek kisebb, teljesen elkülöníthető alstruktúrákból épülnek fel. Ezeket az alstruktúrákat hívjuk doméneknek. A domének több-kevesebb nagyon specifikus szerepű részeket, motívumokat tartalmaznak. Ilyenek például bizonyos anyagok kötőhelyei vagy az enzimek aktívhelyei. A fehérjedoméneket általában többszörös szekvenciaillesztéssel szokták jellemezni. A többszörös illesztésekből származó domén és motívum adatok lehetőséget adnak egy profil létrehozására, amelyek alkalmazhatóak egy fehérje család azonosítására illetve evolúciós kapcsolatok vizsgálatára is. A profilok leírásához könnyen alkalmazható a már említett rejtett Markov model. Ezeket a HMM profilokat tárolva egy géncsalád adatbázishoz jutunk, amilyen például a Sanger Institute PFAM adatbázisa is. A PFAM adatbázis a géncsaládokat leíró profilokat HMM *logo* formában is reprezentálja. (2.12. ábra) Mára már teljes „tudásbázissá” fejlődött, mely tartalmaz annotátorok által karbantartott többszörös illesztéseket, HMM felismerőket, doménleírásokat, keresztreferenciákat a 3D szerkezetekhez, a domént tartalmazó fehérjék „architekturális” leírását, szakirodalmi összefoglalót, …stb.

DOI:10.15774/PPKE.ITK.2014.006



2.12. ábra A PFAM logoval történő reprezentálás egy példája

A képen szereplő PFAM logo részlet a PF00765 azonosító számú, Autoind_synth nevű géncsaládhoz tartozik. Az ábra minden egyes szekvencia pozícióra leírja az adott aminosav előfordulásának valószínűségét: minél nagyobb a betű, annál valószínűbb az előfordulása azon a helyen.

Történeti szempontból érdekes, hogy a fehérjedoméneket először reguláris kifejezésekkel próbálták jellemezni, ez volt az un. PROSITE adatbázis [34], amelyhez a fehérjeszekvenciák motívumainak máig használatos szintaxisát definiálták. A PROSITE kezdte gyűjteni a domének szakirodalmi összefoglalásait is. Ezt az adatbázist ma is fenntartják, de ma már nemcsak reguláris kifejezéseket, hanem profilszerű leírásokat is tartalmaz. Mivel már a kezdeteknél látszott, hogy a reguláris kifejezések nem elég finom leírások, a PROSITE-tal csaknem egy időben megszületett egy másik megközelítés is: az SBASE adatbázisnál használt úgy nevezett doménkönyvtár módszer [35], melyben a doméneket a rájuk jellemző tipikus szekvenciák gyűjteményével jellemezték. Ehhez ugyanis nem kell a nagy emberi munkát követelő többszörös illesztés. Az SBASE az első nyilvánosan hozzáférhető doménszekvencia gyűjtemény volt, később kiegészítették szakirodalmi leírásokkal és statisztikai összegzésekkel, de ma már nem frissítik. A megközelítés előnye, hogy egyszerű szekvenciakeresés révén könnyen megtalálja akár az átlagostól eltérő doménszekvenciákat is, szemben a HMM típusú keresésekkel, amelyek az átlagos doménszekvenciákon teljesítenek a legjobban.

DOI:10.15774/PPKE.ITK.2014.006

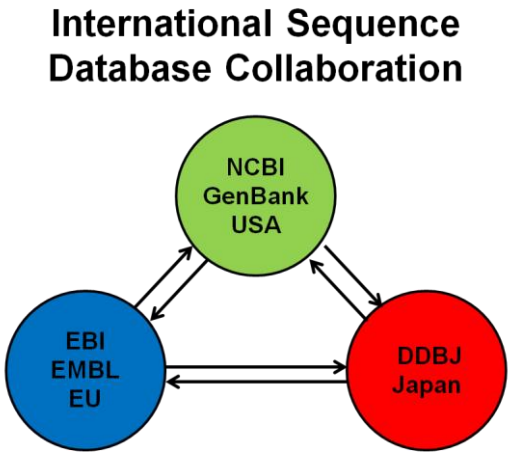
2.5. Bioinformatikai adatbázisok

Dolgozatom elején beszéltem a manapság történő bioinformatikai adatmennyiség robbanásról. Ezt a hatalmas mennyiségű adatot nem elég csupán kinyerni, hanem valahogy tárolni is kell, lehetőleg olyan rendezett formában, amely elősegíti az adatok későbbi elemzését, és az elemzés eredménye hozzá kapcsolható legyen a forrás információhoz.

Napjainkban az adatok bioinformatikai adatbázisokban tárolják. Ezek az információs központok általában egy adattípus tárolására specifikálódnak, ezáltal az adott terület eredményeit a lehető legnagyobb mértékben összefoglalják. A különböző, de összetartozó információk kapcsolatáról az adatbázisok közötti gazdag kereszthivatkozási rendszer gondoskodik. Azokat az adatbázisokat, amelyek magukat forrás adatokat tartalmazzák, *elsődleges* adatbázisnak hívjuk, míg az ezeken az adatokon végzett vizsgálatok eredményeit tartalmazókat *másodlagos* adatbázisnak. A bioinformatikai adatbázisokat általában az általuk tárolt információ típusa alapján csoportosítjuk. A következőekben felsorolom az általam öt legfontosabbnak tartott csoportot.

1) DNS szekvencia adatbázisok

A legfontosabb DNS szekvenciákat tartalmazó elsődleges bioinformatikai adatbázisok a következőek: az NCBI által fenntartott GenBank adatbázis [36], az EBI (European Bioinformatics Institutes) által működtetett EMBL [37] és a japán DDBJ (DNA Database of Japan) [38]. A három adatbázis fejlesztése bár teljesen függetlenül kezdődött el, ma már szoros kapcsolat van közöttük, és egy együttműködés keretében kölcsönösen megosztják egymással az adataikat (2.13. ábra). Az így létrejött óriás adatbázis sajnos redundáns adatokat is tartalmazhat, így az ezeken az adatokon végzett vizsgálatok folyamán erre fokozottan figyelni kell. Mivel én az adatbázis egy kisebb, ellenőrzött részletén dolgoztam (bakteriális teljes genomok), ezért nekem nem kellett számolnom ezzel a hiba lehetőséggel.



2.13. ábra A nemzetközi szekvencia adatbázisok együttműködése

DOI:10.15774/PPKE.ITK.2014.006

2) Fehérje szekvencia adatbázisok

A UniProt Consortium adatbázisa [39] jelenleg a legnagyobb fehérje információ forrás. Több nagyobb adatbázis egyesítésével jött létre, melyek közül a két legfontosabb a TrEMBL és a Swiss-Prot [40]. Az előbbi gépek által annotált szekvenciákat tartalmaz, mely így nagy mennyiségű adatot tartalmaz, de az automatikus eljárások miatt ezek kevésbé megbízhatóak, mint a Swiss-Prot manuálisan annotált és ellenőrzött szekvenciái. Az adatbázis az adott fehérjék szekvenciája és ismert funkcióin kívül egy több szintű klaszterezést is tartalmaz (UniRef), amely a szekvenciák egyezőség szerint csoportosítja az adatbázisban található adatokat. Az adatbázisban tárolt információkon kívül a fehérjékhez egy részletes kereszthivatkozás lista tartozik, amely a fehérjéről összegyűjti szinte minden más bioinformatikai adatbázisban a fontos információkat.

Names and origin

Protein names	*Recommended name:* **Transcriptional activator protein LasR**
Gene names	Name: **lasR** Ordered Locus Names:PA1430
Organism	**Pseudomonas aeruginosa (strain ATCC 15692 / PAO1 / 1C / PRS 101 / LMG 12228)** [Reference proteome] [HAMAP]
Taxonomic identifier	208964 [NCBI]
Taxonomic lineage	Bacteria › Proteobacteria › Gammaproteobacteria › Pseudomonadales › Pseudomonadaceae › Pseudomonas ›

Protein attributes

Sequence length	239 AA.
Sequence status	Complete.
Protein existence	Evidence at protein level

General annotation (Comments)

Function	Transcriptional activator of elastase structural gene (LasB). Binds to the PAI autoinducer.
Miscellaneous	LasR in strain PA103 is not active, this is probably due to the change in position 180 of the sequence.

2.14. ábra Példa egy UniProt rekordra (részlet)

3) Fehérje struktúra adatbázisok

A PDB (Protein Data Bank) [41] a legismertebb biológiai makromolekula struktúra adatbázis, melyet a RCSB (Research Collaboratory for Structural Bioinformatics) tart fent. Az archívum forrása elsősorban röntgen és mágneses magrezonancia vizsgálatok eredményei. Az adatok mind szöveges formában, mind 3D képként is elérhetőek. A projekt elsődleges célja az adatok egységesítése olyan mértékben, amennyire csak lehetséges.

DOI:10.15774/PPKE.ITK.2014.006

4) Bioinformatikai hálózat adatbázisok

A bioinformatika egyik fontos adattípusa a metabolikus útvonalak alkotta hálózatok. Ezeknek az útvonalaknak a legismertebb adatbázisa a japán KEGG (The Kyoto Encyclopedia of Genes and Genomes) [42]. Az adatgyűjtemény főleg molekuláris interakciós hálózatokat, betegség leírásokat és a sejt működéséhez szükséges kémiai vegyületekkel és reakciókkal kapcsolatos adatokat tárol. Az információk kinyerését különböző automatizált kereső algoritmusokkal és többszintű webes eléréssel segíti, így a felhasználó számára megkönnyíti az olykor nagyon bonyolult hálózati rajzok értelmezését. A IX. mellékleten láthatunk egy példát metabolikus útvonal térképre.

5) Bioinfomatikai szöveg adatbázisok

A bioinformatika negyedik alap adattípusa (a szekvencia, a hálózat és 3D struktúra mellett) a szöveg. Ebbe a csoportba tartoznak a könyvek, a folyóiratokban megjelent cikkek és összefoglalók, a konferenciák előadásai és forrás anyagai, bioinformatikai kurzusok tananyagai és gyakorlatilag minden szöveges dokumentum, amely valamilyen módon kötődik ehhez a tudományterülethez. A bioinformatikai folyóirat cikkek legnagyobb gyűjteménye az NCBI által működtetett PubMed adatbázis. Ebben az adatbázisban rengeteg bioinformatikai cikk található meg, olyan formában, amely lehetővé teszi a több szempontú összetett keresések gyors futtatását is. A PubMed a cikkek kiadási adatian kívül tartalmazza azok kivonatait, így a felhasználó rövid betekintést nyerhet a dokumentum tartalmába. Ezenkívül a nyílt elérésű cikkek esetén a teljes szöveg is elérhető és a megtekintéskor a felhasználónak lehetősége van több szöveges formátum közül is választania.

2.6. Fontosabb funkció adatbázisok

A funkciók leírásaira kétféle származtatott adatbázist használnak. Az első típus a funkció szerint csoportosított fehérjeszekvenciák gyűjteménye, melynek alaptípusa a COG. [43] A másik típus a funkciók szabványos leírására koncentrál, amelyet fogalmi hierarchiákban, szabályokban, szaknyelvi nevén ontológiákban foglalnak össze. Ennek alaptípusa a GO, melyet a *Gene Ontology Consortium* fejleszt folyamatosan. [44] A helyzetet bonyolítja, hogy a fehérjefunkciók leírásához már a COG készítői is kifejlesztették a maguk ontológiáját, ami eltér a GO leírásoktól. Ebben a fejezetben a COG és a GO megközelítését ismertetem.

DOI:10.15774/PPKE.ITK.2014.006

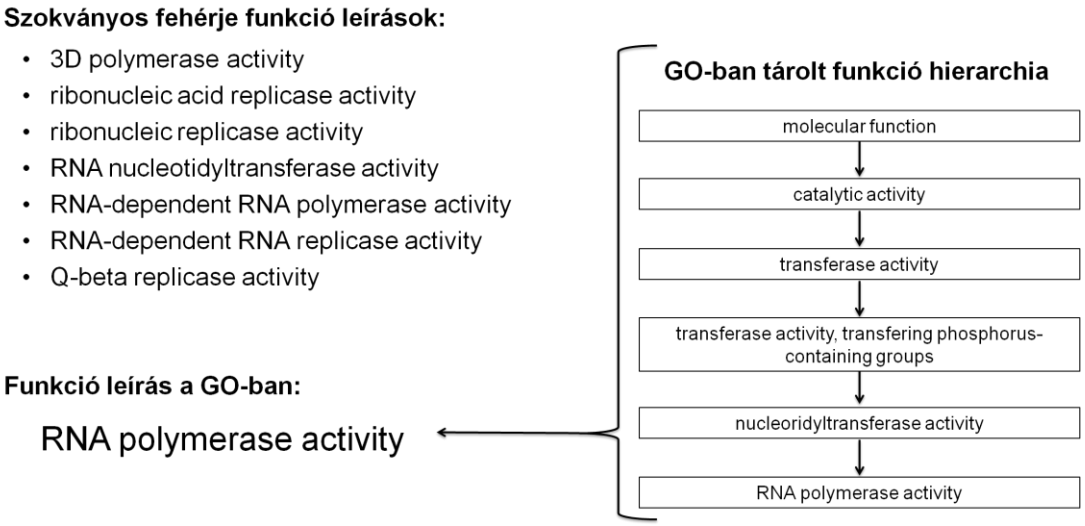
2.6.1. Clusters of Orthologous Groups

A COG [43] (*Clusters of Orthologous Group* = Ortológ csoportok klaszterei) az egyik legismertebb fehérje-adatbázis, amely a genomok filogenetikai osztályozásán alapul. Mindegyik COG csoport valószínűsíthető ortológokat tartalmaz, figyelembe véve az evolúciós kapcsolatokat, mint például a génduplikáció. Az adatbázist eredetileg kísérleti céllal hozták létre 2000-ben, hogy az újonnan szekvenált gének funkcionális annotációját megkönnyítse. A korai tapasztalatok alapján 17 főbb funkcióosztályt hoztak létre, amelyek között alapvető sejtfunkciók és biokémiai aktivitások mellett a még ismeretlen funkciók is helyet kaptak. A csoportok kapcsolatrendszere jól felépített, így az adatbázis könnyen frissíthető az újonnan megismert génfunkciókkal. Bár az adatbázis nagy reményekkel és jelentős sikerekkel indult, pár évvel a megalkotása után a frissítése befejeződött. A fejlesztése alatt összesen 4872 különböző csoportot azonosítottak, melynek körülbelül egy negyede (1346) tartozik az ismeretlen funkciójú gének csoportjába.[45] A COG adatbázist mindezzel együtt szinte kötelezően használják a baktériumgenomok annotációjához. Pótlására, frissítésére készült pl. az EGGNOG adatbázis, amelyik a COG emberi tudással definiált csoportjai mellett automatikusan generált csoportokat is tartalmaz.

2.6.2. Gene Ontology

Az előzőekben láthattunk lehetőségeket a gének annotálására, és a biológiai folyamatokban betöltött szerepének megismerésére. Azonban ezt a szerepet nem elég felismerni, hanem valamilyen rövid formában le is kell tudnunk írni. Ez elsőre nem tűnik bonyolult feladatnak, de mivel egy tulajdonság vagy szerepkör többféleképpen is megfogalmazható, illetve több funkció esetén különböző fontosságot adhatunk az egyes szerepeknek, ezért fontos a leírás egyértelműsége. Ezt a célt szolgálja a GO (*Gene Ontology*). [44] Ez az adatbázis a gén tulajdonságokat három csoportba osztja: biológiai folyamat, molekuláris funkció és sejt komponens. A csoportokba tartozó ontológiák gráfként vannak ábrázolva, ahol minden alacsonyabb szint egyre specifikusabb leírást jelöl. A *Gene Ontologia* használatával elkerülhető a korábbi káosz, ami a nem egyértelmű funkció nevekből adódott. Bár mindegyik leírja az adott funkciót, automata keresés esetén nehéz eldönteni ezek azonosságát. (2.15. ábra)

DOI:10.15774/PPKE.ITK.2014.006



2.15. ábra Példa a Gene Ontology egy osztályozására

Bal oldalt látható egy adott fehérjecsalád funkciójának különböző elnevezései, és ugyan ennek a funkciónak a *Gene Ontology* által adott terminusa. A témához értő személy könnyen megállapíthatja, hogy a funkció leírások ugyanazt a biológiai szerepet takarják, de egy automatikus keresés esetén az algoritmus nem tudja ezt eldönteni. A *Gene Ontology* által adott kifejezés és a hozzá tartozó hierarchikus leírás azonban megkönnyíti az automatizált eljárások feladatát. Jobb oldalt látható az adott terminus hierarchiája.

DOI:10.15774/PPKE.ITK.2014.006

3. Célkitűzések

Dolgozatom konkrét célja a baktériumok egyik legfontosabb jelzőrendszerét, az úgy nevezett *quorum sensing* rendszert alkotó gének felmérése volt a ma hozzáférhető nyilvános szekvencia-adatbázisok felhasználásával. Ehhez a munkához az egyik legelterjedtebb ilyen rendszert, a Gram-negatív baktériumok úgy nevezett AHL jelzőrendszerének génjeit választottam ki tárgyként, mert ez jól definiált, részleteiben is ismert rendszer, ugyanakkor sok változata ismeretes.

Munkatervem célja kettős volt: egyrészt le akartam írni a bakteriális kommunikáció génjeinek szerveződését, másrészt egy olyan automatizálható rendszer kialakítása, amellyel a hasonló génszerveződések is leírhatók. Ezeknek az alapoknak figyelembevételével a következő célokat tűztem ki:

1. A Gram-negatív baktériumok *quorum sensing* génjeinek megismerése, és a szerveződésük leírására alkalmas jelölési rendszer kialakítása. A jelölésrendszernek alkalmasnak kell lennie a későbbi esetleges kibővítésre, és más kis elemszámú alrendszerek felírására is.
2. A *quorum sensing* gének felmérésére alkalmas számítógépes eljárás kidolgozása, mely minimális felhasználói beavatkozás mellett a lehető legbiztosabb eredményt adja. A helyesség biztosítására többszintű ellenőrzési módszer létrehozása.
3. A *quorum sensing* gének felmérése a publikus adatbázisokban. Az eredmény helyességének nyomon követése a teljes folyamat során. A talált gének és legfontosabb tulajdonságainak tárolása és különböző szempontok alapján történő csoportosítása, mely lehetővé teszi az adatok későbbi elemzését.
4. A talált gének bemutatására alkalmas weboldal kidolgozása, mely az informatikában kevésbé jártas felhasználók számára is könnyen átlátható és használható.

DOI:10.15774/PPKE.ITK.2014.006

4. Adatok és módszerek

4.1. Az adatok forrása

A munkámhoz nyilvános adatbázisokban fellelhető szekvenciális adatokkal dolgoztam. Az egységesség kedvéért kizárólag az egyesült államokbeli National Institutes of Health (NIH) National Center for Biotechnology Information (NCBI) adatait használtam. [46] Innen a bakteriális adatok háromféle adattípusát töltöttem le: a teljes genomokat, a részletes genomokat és az egyéni DNS szekvenciákat.

Az adatbázisban szereplő adatokat kétféle módszerrel értem el: ha kevés információra volt szükségem egy kézzel végzet vizsgálathoz vagy ellenőrzéshez, akkor a hivatalos honlapon lévő összetett kereső segítségével kerestem meg és töltöttem le a szükséges adatokat. Az automatizált algoritmusok azonban az NCBI FTP szerverét [47] használták, amely megkönnyítette a nagy mennyiségű adatok letöltését.

Az adatokhoz való hozzáférésnél két eltérő szemléletmódot alkalmaztam: offline és online. Offline esetben az adatbázis összes szükséges adatfájlját előre letöltöttem a gépre, és a későbbiekben ebből nyertem ki a szükséges információt. Ennek a módszernek az előnye, hogy nem igényel aktív internet kapcsolatot a letöltés után, a merevlemezen tárolt adatbázis elérése sokkal gyorsabb és a több órákig futó algoritmusok nincsenek kitéve a túlterhelt adatbázisok elérési kimaradásainak. Hátránya viszont hogy az egész adatbázist tárolni kell, melynek mérete csak legszükségesebb adatbázis részletek esetén is több 10 GB, melynek legnagyobb részét nem is használjuk. Online adatbázis hozzáférés esetén az automatizált algoritmus egyesével tölti le a szükséges adatbázis fájlokat, ezáltal mindig a legfrissebb adatokon dolgozik, és csak a tényleg szükséges fájlok kerülnek letöltésre. Például a bakteriális *ptt* formátumú fájlok esetén csupán a teljes adatbázis kevesebb, mint 10% került vizsgálat alá.

A munkám során először offline adatbázis elérést használtam, de mivel a teljes automatizálás megköveteli az adatok frissességét, ezért átértem az online módszerre, ami ugyan lassabb, de ez a megoldás közelebb áll a megoldandó problémához. Így az utolsó teljes adatbázis letöltés 2012. januári, de azóta a keresési eredmény többször frissítve lett az új algoritmusfutások által.

Habár kizárólag az NCBI adatbázis adatait használtam az algoritmus futása során, a fejlesztés során többször végeztem kézi ellenőrzéseket a bizonytalan találatok esetén. Ezek során más adatbázisokat is használtam, többek között a UniProt [48] adatbázis UniRef klasztereit, amelyek segítettek ellenőrizni az eredményeket: ha a találatok néhány klaszter segítségével lefedhetőek, az megerősíti a helyességüket. A klaszterek kimaradt tagjai pedig segítenek javítani, tovább fejleszteni a keresés alapját képező HMM profilokat.

DOI:10.15774/PPKE.ITK.2014.006

4.2. A keretrendszer kialakítása

Mielőtt nekiállunk elkészíteni egy automatizált keresési rendszert, mindenképp meg kell határoznunk, hogy milyen adat és programozási környezetben fogjuk használni. Ennek megtervezése minden szoftver esetén fontos, mert a későbbi változtatás rendkívül időigényes lehet és nem is biztos, hogy a követelmények szinten tartásával lehetséges. A projektemnél a folyamat során kizárólag szöveges fájlokat használtam, csak bizonyos eredmények kézi elemzésénél volt szükségem képfájlokra (hasonlósági fák, kladogramok), viszont ezeknek létezik egzakt szöveges leírása is, így egyféle adattípussal kellett foglalkoznom: szöveges fájllal. Mivel a szöveges fájlok karakterkódolása is eltérhet különböző rendszereken, ezért a minden platform által támogatott, BOM-nélküli UTF8 kódolást választottam.

Mivel az adatok kezelése nem befolyásolja a keretrendszer kiválasztását, ezért csak a használni kívánt programcsomagok és szoftverek támogatását kellett figyelembe vennem. Első tervezési lépés az operációs rendszer kiválasztása volt. Mivel a bioinformatikai alkalmazások döntő többsége Unix alapú operációs rendszerre van optimalizálva, ezért alap rendszernek a Linuxot választottam. Bár legtöbb esetben kipróbáltam a Windows környezetben futó alternatívákat, tapasztalatom szerint ezeket használata nehézkesebb, és általában lassabban is futottak. A munka során készült programszkriptek azonban az esetek döntő többségében nem használták ki a Unix operációs rendszer speciális lehetőségeit, így az általam létrehozott program apró módosítások után más rendszeren (Windows) is futtatható volt.

Az operációs rendszer kiválasztása után a programozási nyelvet kellet kiválasztanom. Ez a legfontosabb tervezési lépés, mert a használt nyelv döntően befolyásolja a program felépítését, hatékonyságát és nem utolsósorban fejlesztési nehézségét. Nincs „tökéletes” választás, mivel sok különböző alfeladatot kell elvégezni, és egy adott nyelv egyes részleteket gyorsan és hatékonyan old meg, míg másokat lassan és nagy erőforrás igénnyel. Erre megoldás lehet ugyan több nyelv közös használata, de ez ennyire kis projektnél nem hozna akkor javulást, mint amennyit a programrészletek összehangolása elvenne. Így egy nyelv választása mellett döntöttem. Mivel nagyobbrészt fájlfeldolgozás, külsőparancs hívás és szövegmanipulálás szükséges a feladatok megoldásához, ezért egy szkriptnyelvet volt érdemes választanom, mivel a nyelvek ezen típusa alacsony erőforrás igényű. A bioinformatikában két ilyen nyelv használata terjedt el: a Perl és a Python. Mindkettő megfelelő módon támogatja a reguláris kifejezések használatát, az egyszerű fájlbeolvasást és mindkettővel viszonylag gyors kód írható. Bioinformatikai modul is található hozzájuk (BioPerl[49, 50] és BioPython[51]), és rengeteg hasznos példa és segédlet fellelhető az interneten. Eleinte a Perl nyelvet használtam, mert az elemi lépések elkészítéséhez jobbnak bizonyult az eszközkészlete. Amikor azonban a teljes munkamenet összeillesztésére, és az automatizáláshoz nélkülözhetetlen folyamatos adat validálásra került a sor, a Python nyelv nyújtott megbízhatóbb teljesítményt, így végül a végső

DOI:10.15774/PPKE.ITK.2014.006

program ezen a nyelven készült el. Bár a két nyelv szintaktikája elégé eltérő, még sem okozott nagy problémát a nyelvek közötti átállás, mivel a szemantikájuk nagyon hasonló, így az algoritmusok döntő többségét nem kellett áttervezni. A szkriptnyelveknek a hatékony szövegfeldolgozáson kívül van egy közös tulajdonságuk: rendkívül robosztusak. Ez azt jelenti, hogy futásközben fellépő hiba esetén is tovább folytatja a futást (kivéve pár kritikus hiba esetén). Ez azért hasznos tulajdonság, mert egy több tízezer soros fájl feldolgozása nem szakad meg pár sor hibája miatt, így csak a hibás sorok esetén kell az algoritmust újra futtatni. Ez viszont megköveteli a program futása közben történő folyamatos eredmény ellenőrzést, mert ha egy olyan adat lesz hibás, amit a későbbiekben még felhasználunk, akkor ez a hiba tovább gyűrűzik az egész folyamat során, és elronthatja a végeredményt.

A munkamenet során több bioinformatikai szoftvert is használtam, ezek közül a fontosabbak a következők:

ClustalW	2.0.11.-es verzió	(2.3.1 fejezet)
hmmer	3.0.-ás verzió	(2.3.2 fejezet)
BLAST	2.2.25.-ös verzió	(2.3.3 fejezet)
Artemis	15.0.0.-ás verzió	(5.1.2 fejezet)
PHYLIP	3.695-ös verzió	(4.3.1 fejezet)
Jalview	2.0.1.-es verzió.	(I. Melléklet)

Az algoritmus tesztelése egy Intel® Core™ i7-2630QM processzorral rendelkező gépen történt, 5400 rpm sebességű merevlemez eléréssel.

4.3. Hasonlósági fák

Az adatok megjelenítésére hasonlósági faépítő algoritmusokat használtam. A hasonlósági fa egy gráfelméleti binárisfa, amelynek a levelei tartalmazzák a szekvenciákat. A fában a levelek közötti távolság pedig egyenesen arányos a szekvenciák hasonlóságának mértékével. A szekvenciák homológijának vizsgálatára több módszer is van, melyek nagyban befolyásolják az elkészített fa tulajdonságait és pontosságát. Egy biológiai szemmel kifogástalan fa elkészítése rendkívül bonyolult feladat. Mivel azonban én csupán a szignifikáns csoportosulások megkeresésére használtam a módszert, így számomra a kevésbé pontos megoldások is megfeleltek. Ennek ellenére a fa készítésének legalapvetőbb szabályát mindenképp szem előtt kell tartani: csak ténylegesen hasonló szekvenciákból készítsünk hasonlósági fát, mert különben a fa elveszti az információ tartalmát. Ezért minden esetben csak azonos típusú gének szekvenciából készítettem fát, csak különböző feltételeknek megfelelően válogattam ki őket. (Például: a *burkholderiák* rendjébe tartozó baktériumok *luxR* génjei vagy a R2 típusú topológia *luxR* génjei.)

DOI:10.15774/PPKE.ITK.2014.006

4.3.1. Fakészítési algoritmusok

Mivel a rendelkezésre álló DNS szekvenciákról nem rendelkeztem evolúciós adatokkal, így a hasonlósági fa építéséhez távolság alapú eljárásokat használtam, ahol a távolság a szekvenciák páros illesztéséből származó szerkesztési (Levenshtein) távolság. A távolságmátrixból több eljárás segítségével és elkészíthetjük a filogenetikus fánkat:

UPGMA: egyszerű agglomerációs vagy hierarchikus klaszterező algoritmus, mely teljesen különálló levelekből indulva, minden lépésben a két legközelebbi részfát összekapcsolva hozza létre a végső fát.

Neighbor-joining: mohó algoritmus, ami egy csillag alakú fából kiindulva, minden lépésben annak a két ágnak az összevonását végzi el, aminek hatására az össz ághossz a legkevesebb lesz.

Bár egyik távolság alapú algoritmus sem tud annyira pontos megoldást adni, mint a modell alapú módszerek, a számítás igényük sokkal kisebb, így nagyobb adathalmaz esetén is alkalmazhatóak voltak. A második metódus segítségével pontosabb eredmény érhető el, így munkám során ezzel dolgoztam. Ezeket az algoritmusokat elsősorban a *PHYLIP* [52] programcsomagon keresztül használtam, ami lehetővé tette a műveletek pontos paraméterezését. Emellett alkalmaztam a *ClustalW* algoritmus által készített irányító fát predikciós célokra.

DOI:10.15774/PPKE.ITK.2014.006

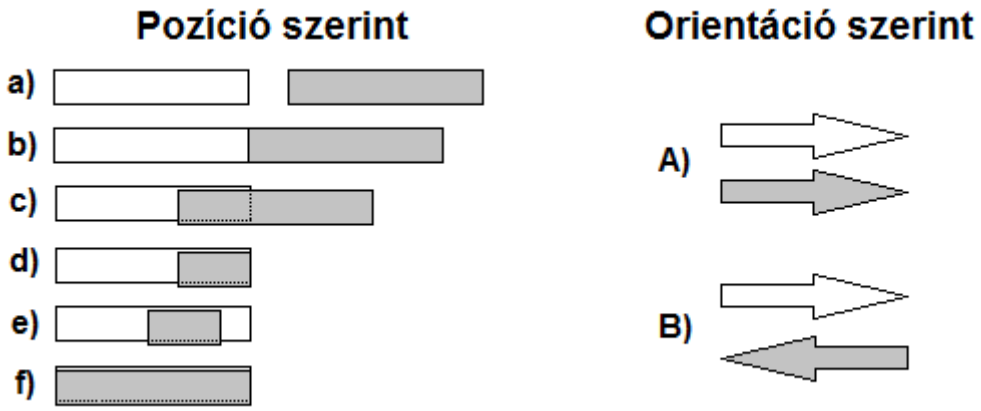
5. Eredmények I.

5.1. A munkamenet megtervezése és a program kidolgozása

5.1.1. Használt jelölések

A *quorum sensing* gének elhelyezkedését két szempont alapján vizsgáltam. Az első szempont, hogy a gének mennyire fedik át egymást. Ez az adat a gének kezdő és végpozíciójának összevetésével tudható meg. A különböző eseteket az 5.1. ábra mutatja. Az egyes esetek nevei:

a) diszjunkt b) érintkező c) részlegesen átfedő
d) közös végpont e) tartalmazó f) azonos szegmens

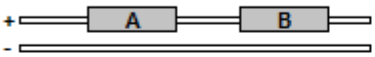
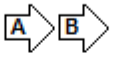
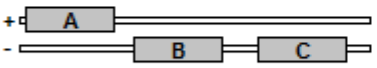
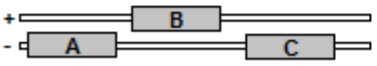
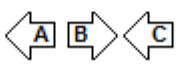


5.1. ábra Lehetséges pozíciók és orientációk két gén esetén.

A másik vizsgálati szempont a gének orientációja. Ez az irányultság a gén kifejeződésének irányával van összefüggésben, ami attól függ, hogy melyik DNS szálon található. Az általam használt konvenció a topológiák szemléltetésekor: a + szál egy jobbra mutató nyíllal van reprezentálva, a – szál egy balra mutató nyíllal. Ez alapján két gén egymáshoz viszonyított orientáltsága két féle lehet: egy irányba mutató (parallel, A) vagy különböző irányba mutató (anti-parallel, B). (5.1. ábra)

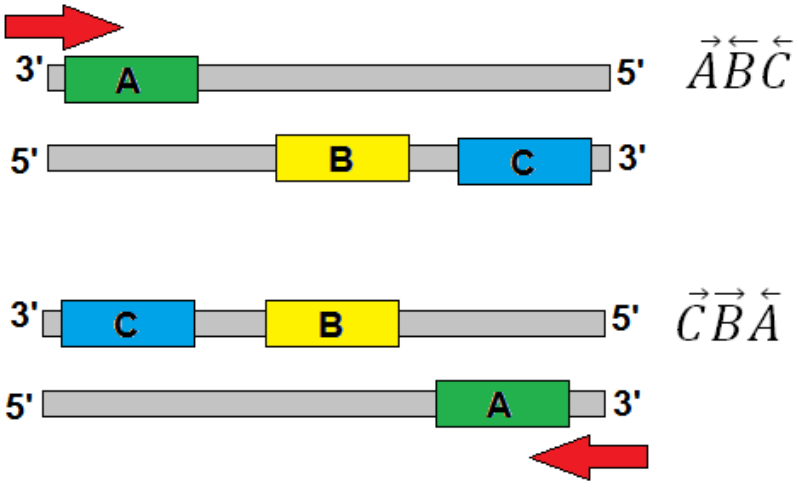
Mielőtt megvizsgáltam a topológiák génjeinek egymáshoz viszonyított orientáltságát, érdemes volt meghatároznom egy egyszerűsített, formális felírást, amivel jelölni tudtam az adott elrendeződést. Én a következő jelölést használtam a topológiák leírására: felsorolom a géneket a DNS-en való pozíció sorrendjébe, majd utána a gének fölé rajzolt nyíllal jelzem az irányultságot. Ez a jelölés egyszerű és reprezentatív, viszont egyszerű szöveges fájlokban történő leírásra nem használható. Ilyen esetekben az orientáltságot nem nyilakkal jelöltem, hanem a géneket jelölő betűk után a megfelelő sorrendben felsoroltam a szálak jelét. Példák a nyíllal és a DNS-en való ábrázolással a következő képen láthatóak. (5.2. ábra)

DOI:10.15774/PPKE.ITK.2014.006

DNS	Nyílakkal	Felírás	Cleartext Felírás
		$\vec{A}\vec{B}$	AB++
		$\vec{A}\overleftarrow{B}\overleftarrow{C}$	ABC+--
		$\overleftarrow{A}\vec{B}\overleftarrow{C}$	ABC-+-

5.2. ábra Példák a topológia felíró jelölésre.

A DNS szálak szimmetriájának 3'→5' és 5'→3' következménye az, hogy bizonyos topológiák lényegében ugyan azok. Mivel nincs konkrétan szabályozott meghatározás arra, hogy melyik DNS szál a pozitív (+) és melyik a negatív (-), ezért azokat a topológiákat azonosnak tekintettem, ahol a DNS megfordításával ugyanazt az elrendeződést kapjuk. Mint például $\vec{A}\overleftarrow{B}\overleftarrow{C}$ és $\vec{C}\vec{B}\overleftarrow{A}$ esetében. A szimmetriát az 5.3. ábra reprezentálja, ahol jól látszik, hogy ha a piros nyilak mentén nézzük a DNS-en a gének elhelyezkedését (mindegyik 3' oldal), ugyanazt fogjuk látni. A szimmetrikus topológiák esetén azonban el kell döntenünk, hogy a két lehetőség közül melyik névvel illetjük az adott topológiát. Például lehetőségünk van a topológiák ábécé sorrendben vett változatai közül az első kiválasztani, vagy kikötjük, hogy az első gén mindig a pozitív szálon legyen, esetleg a topológiában részt vevő gének között felállítunk egy saját relációt. Munkám során én az utolsó lehetőséget választottam: a *luxR* gén rendelkezett a legnagyobb prioritással, a *luxI* gén pedig a legkisebbel. A választásomnak semmilyen biológiai alapja sincs, csupán számomra esztétikusabban tűntek az **R**-rel kezdődő topológiai felírások.



5.3. ábra A topológiák szimmetriájának szemléltetése.

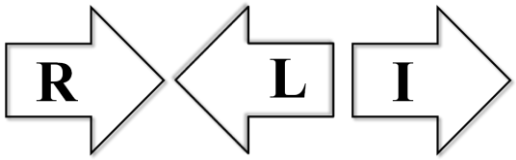
Mivel a két DNS szálat nem különböztetjük meg egymástól, így a képen szereplő két topológia ($\vec{A}\overleftarrow{B}\overleftarrow{C}$ és $\vec{C}\vec{B}\overleftarrow{A}$) ugyanannak tekinthető.

DOI:10.15774/PPKE.ITK.2014.006

5.1.2. Topológiák szemléltetése

Egyszerű ábrázolás

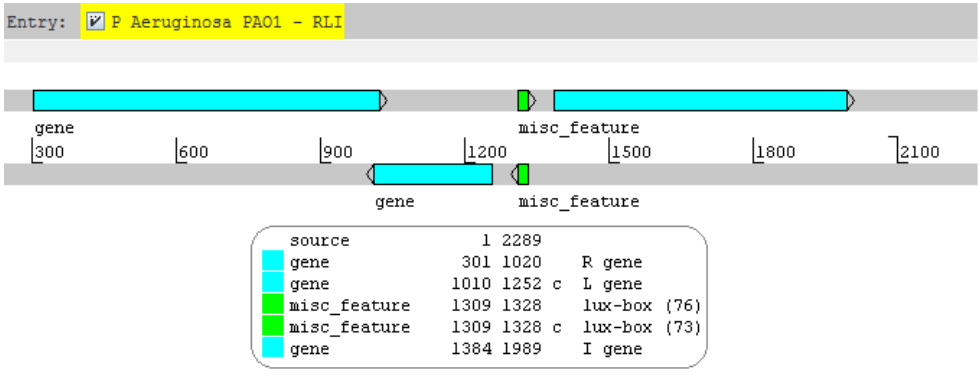
A topológiák szemléltetésének legegyszerűbb formája, amikor a géneket egyszerű, azonos méretű nyíllal ábrázoljuk. Ebben az esetben a gének egymáshoz viszonyított helyzete és iránya a hangsúlyos. A könnyű áttekinthetőség érdekében figyelmen kívül hagyjuk a gének méreteit, átfedésüket és minden egyéb tényezőt. Ez az ábrázolás az elrendeződések keresésének elején a leghasznosabb, mivel gyorsan kapunk egy könnyen elemezhető ábrát az adott topológiáról. Elkészítése nem igényel specifikus programot, így szinte bármilyen szoftverkörnyezetben elkészíthető. Weblapon való dinamikus megjelenítéséhez nem szükséges külön erőforrás, gyakran használt eszközökkel megoldható. Egy $\vec{R}\overleftarrow{L}\vec{I}$ típusú topológia egyszerű ábrázolását mutatja az 5.4. ábra.



5.4. ábra Példa topológiák egyszerű ábrázolására

Artemis

Az egyszerű ábrázolásnál összetettebb képet kapunk a Sanger Institute *Artemis* [53] nevű genomábrázoló és annotációs eszköze segítségével. [54] Ez a szemléltetési módszer lehetőséget ad a gének méretének és átfedésének megjelenítésére is. Az orientáltságbeli különbséget a különböző irányú nyilakon kívül a különböző DNS szálakra rajzolás is kiemeli, így az azonos transzlációs irányú gének könnyebben vizsgálhatóak. Az *Artemis* szoftver eszköztára lehetőséget ad a géneken kívül más DNS elemek megjelenítésére is. A *Pseudomonas Aeruginosa PAO1* egy $\vec{R}\overleftarrow{L}\vec{I}$ típusú topológiájának *Artemis*szel történt szemléltetését mutatja az 5.5. ábra. A kommunikációért felelős géneken kívül (*luxI*, *luxR* és *rsaL*) megjelenítésre került a DNS egy *lux-box* nevű, *quorum sensing*gel kapcsolatos jellegzetessége is.

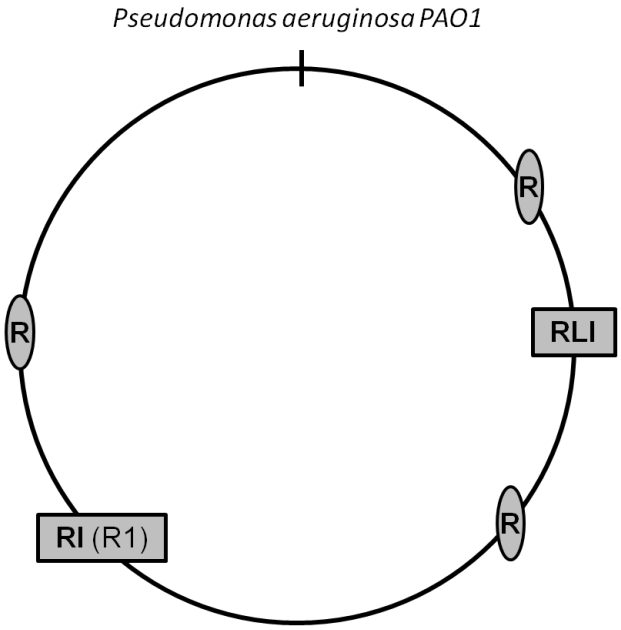


5.5. ábra Artemis segítségével történő topológia ábrázolás egy példája

DOI:10.15774/PPKE.ITK.2014.006

Kromoszóma térkép

A kromoszóma térkép az eddigi ábrázolás módoktól eltérően nem egy elrendeződésben szereplő gének egymáshoz való viszonyát hivatott reprezentálni, hanem az egy kromoszómán lévő topológiák elhelyezkedését a genomban. Ennek segítségével nem csak a topológiák típusa és száma alapján tudjuk összehasonlítani a baktériumokat, hanem konzervált elhelyezkedési mintákat is kereshetünk. Az 5.6. ábra a *P. Aeruginosa PAO1* baktérium genomját reprezentálja. Mivel a legtöbb baktériumnak a DNS-e cirkuláris, ezért választanunk kell egy viszonyítási pontot, ami alapján elhelyezzük a géneket a rajzon. Jelen esetben én a GenBank rekord vágási pontját választottam, és rögzítettem a kör legfelső pontján. Egy genom vizsgálata esetén ennek nincs jelentősége, mivel a térkép bármekkora mértékbe elforgatható. Ha viszont több genomot szeretnénk összehasonlítani, akkor mindenképp szükséges egy biológiai értelemmel bíró közös viszonyítási pontot választani. Erre tökéletesen alkalmas egy mindegyik genomban szereplő, jól ismert gén.



5.6. ábra Példa a topológiák kromoszóma térképére

5.1.3. Szekvenciák hasonlóságának ábrázolása

A topológiák összehasonlító vizsgálatának egyik része az elrendeződéseket alkotó gének különálló vizsgálata is. Ez által megtudhatjuk, hogy az azonos típusú topológiákban szereplő gének mennyire hasonlítanak egymásra, és ezeknek a hasonlóságoknak van-e kapcsolata a baktériumok evolúciós vagy taxonómiai rokonságával. Génszekvenciák hasonlóságának ábrázolására legjobb módszer a filogenetikus fa készítése.

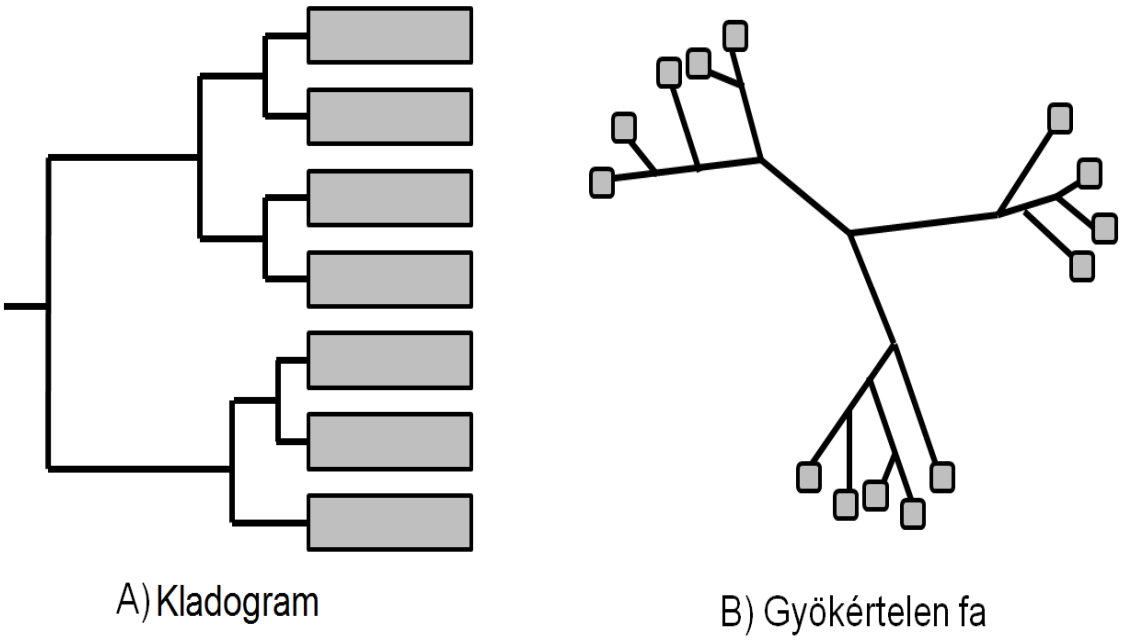
DOI:10.15774/PPKE.ITK.2014.006

A fa megjelenítése

A fakészítő algoritmusok legtöbbje az eredményt Newick formátumban adják meg. Ez a formátum a fa szerkezetét zárójelezéssel kódolja. Ezenfelül - amennyiben szükséges - eltárolja a fa ágainak hosszát is. Az elkészült fákat – függetlenül a készítés algoritmusától – különböző módon tudjuk ábrázolni attól függően, hogy a fa milyen célból készült: adatokat akarunk leolvasni róla (csoportok meghatározása), vagy csupán szekvenciák hasonlóságának szemléltetése a cél. Az adatok függvényében én két különböző ábrázolási technikát használtam. (5.7. ábra)

A) Kladogram: Gyökérrel rendelkező fa (irányított gráf), ami egyértelműen azonosítja legközelebbi közös őst (Egy adott szekvencia, ami nem szerepel a bemenetben). A bemeneti szekvenciák a fa leveleiben szerepelnek és a gyökértől való távolságot a hipotetikus közös őstől való genetikai távolság határozza meg. A gyökér azonosításához általában egy olyan szekvenciát használnak, ami vizsgált csoporttal csak távoli kapcsolatban áll.

B) Gyökértelen fa esetén a bementi szekvenciák távolságát és kapcsolatait ábrázoljuk anélkül, hogy bármit feltételeznénk a származásukról. Egy gyökeres fából minden esetben készíthetünk egy gyöktelent, viszont gyökértelen fa esetén rendszerint nem lehetséges gyökér elhelyezése további információk hozzáadása nélkül.



5.7. ábra Az általam használt két hasonlósági fa típus

DOI:10.15774/PPKE.ITK.2014.006

5.1.4. Szekvencia illesztések konzerváltsága

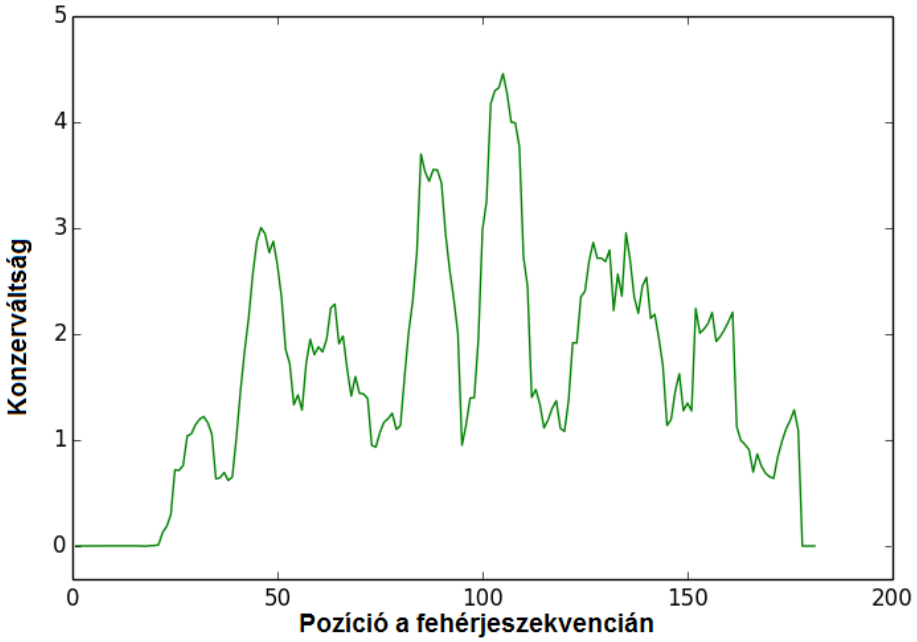
A szekvenciák hasonlóságának feltárásához nyújt lehetőséget a többszörös illesztésük vizsgálata. Az illesztés egy oszlopában az aminosavak minél kevésbé változatosak, az adott pontban annál *konzervált*abb az illesztés. Ennek a tulajdonságnak a számszerűsítésével lehetőségünk van illesztés legmeghatározóbb régióinak meghatározásához. Ezek a szekvencia szakaszok segítségükre lehetnek egy gén különböző típusainak hasonlóság vizsgálatában.

A *konzerváltság*ot több féle módszerrel is lehet mérni, mint például a leggyakrabban előforduló aminosav számossága. Ez azonban nem ad elég árnyalt eredményt és nem veszi figyelembe a különböző aminosavak hasonlóságát sem. Ezért az *Emboss* programcsomag *plotcon* nevű ábrázolóprogramjához hasonló elvet használtam: az adott oszlopban szereplő aminosavakat minden lehetséges módon párba állítottam, majd *helyettesítési mátrix* (*Blosum62*) segítségével megnéztem mennyire hasonlítanak egymásra. Ezután ezeknek az értékeknek vettem az átlagát. Az így kapott érték jól reprezentálja az adott oszlop *konzerváltság*át. Az értékeket a szekvencián való helyzetüknek megfelelően grafikonon ábrázolva egy konzerváltságot jól szemléltető ábrát kapunk. Mivel ez a módszer érzékeny a kis különbségekre ezért a görbét megfelelő ablak használatával lehet simítani.

A *konzerváltság* kiszámításának képlete:

$$Conservation(W) = \frac{\sum_{x,y \in W} \boldsymbol{M}(x,y)}{S} \qquad ahol \quad S = \frac{ws * n * (n-1)}{2}$$

A képletben szereplő *W* az ablak oszlopaiban szereplő összes aminosavat tartalmazó halmaz, *x* és *y* a *W* halmaz egy-egy eleme, az ***M*** függvény a *helyettesítési mátrix* értékét visszaadó függvény, *ws* az ablak mérete és *n* az illesztésben szereplő szekvenciák száma.



5.8. ábra Az rsaM gének konzerváltságát bemutató grafikon

DOI:10.15774/PPKE.ITK.2014.006

5.2. Munkafolyamat lépései

5.2.1. HMM Profil és adatbázisok

A feladat megoldásának nulladik lépése, hogy az automatizált futáshoz szükséges összes fájlt előkészítjük, illetve elkészítjük. Ez a művelet nem automatizált, mivel az információk nagyobb része kizárólag a felhasználótól függ. A legalapvetőbb dolog a fehérjecsaládok kiválasztása, amelyeknek elhelyezkedését vizsgálni akarjuk. Ezeket a fehérjecsaládokat egy konkatenált (összefűzött) HMM profil segítségével írjuk le. Ha a HMM profil készen van, akkor a keresés során használt többi változót tartalmazó fájlt kell elkészítenünk. Ez az információs fájl tartalmazza a fehérjecsalád rövidített nevét és a keresés azon feltételeit, amely alapján elfogadunk egy találatot a fehérjecsalád tagjának vagy sem. Ezek a feltételek határozzák meg a végső eredmény helyességét és a későbbi felhasználás lehetőségeit; ha engedékeny szabályokat használunk, az eredmények halmaza nagyobb lesz, így nagyobb eséllyel tartalmaz minden szükséges fehérjét, de megbízhatósága kisebb lesz, emiatt az adatok további vizsgálata és ellenőrzése szükséges.

Miután a fehérjecsalád leírásával végeztünk, ki kell választanunk az adathalmazt, amelyen futtatni szeretnénk a keresést. Alapértelmezés szerint az NCBI GenBank összes bakteriális adatát tartalmazó adatbázisán fut a keresés, melyet online vizsgál, de lehetőség van a felhasznált adathalmaz szűkítésére is, így nem kell kivárni a teljes kereséshez szükséges időt, ha például csak a *Pseudomonas* nemzetségbe tartozó baktériumokat akarjuk megvizsgálni. A keresési eljárás során mind a keresés dátuma, mind az adathalmaz neve eltárolásra kerül az eredmény fájl fejlécében, hogy a későbbiekben könnyen azonosítható legyen, hogy milyen keresési eredmény található az adatfájlban, és mennyire friss adatokon futott.

5.2.2. HMM keresés futtatása

A munkamenet első lépése a HMM profil alapján történő keresés futtatása. Ez a megadott HMM profilokkal történik a beállított adathalmazon. Ebben a lépésben az adatok értékelése még nem történik meg, csupán az egyesével letöltött FASTA formátumú adatbázisfájlokon lefut a *hmmer* [55] programcsomag *hmmsearch* algoritmusa, az eredményeket pedig egy szöveges fájlba gyűjti össze a program. A fájlban a következő információkat tároljuk el a génekről: génazonosító, a HMM profil melyik doménjével van találat, ennek a találatnak a valószínűsége (*e-value*) és a forrásfájl neve (amely tartalmazza a baktérium nevét és a kromoszóma azonosítóját, így a futás későbbi lépéseiben könnyen elérhetjük az adott baktérium különböző adatfájljait is). A fájl első pár sorában egy fejléc foglal helyet, amely tartalmazza a keresés fontosabb paramétereit, mint például a vizsgált

DOI:10.15774/PPKE.ITK.2014.006

baktériumok száma, a lefutatott HMM keresések száma (ez egyenlő a szűkített adatbázisban található fasta fájlok számával) és a keresés találatainak száma. Ez a szöveges fájl adódik tovább az algoritmus következő lépésének.

Az egyesével történő letöltés lassítja ugyan az algoritmus futását, viszont biztosítja, hogy minden esetben automatikusan a legfrissebb adatbázist éjük el, és nem kell a merevlemezen tárolni viszonylag nagy mérettel rendelkező adatbázis archívokat. A keresést végrehajtó program részlet azonban modulárisan lett megírva, így amennyiben szükséges, az adatfájl letöltő függvény könnyedén kicserélhető egy olyannal, ami egy a számítógépen található könyvtárból vagy archívumból veszi az adatokat.

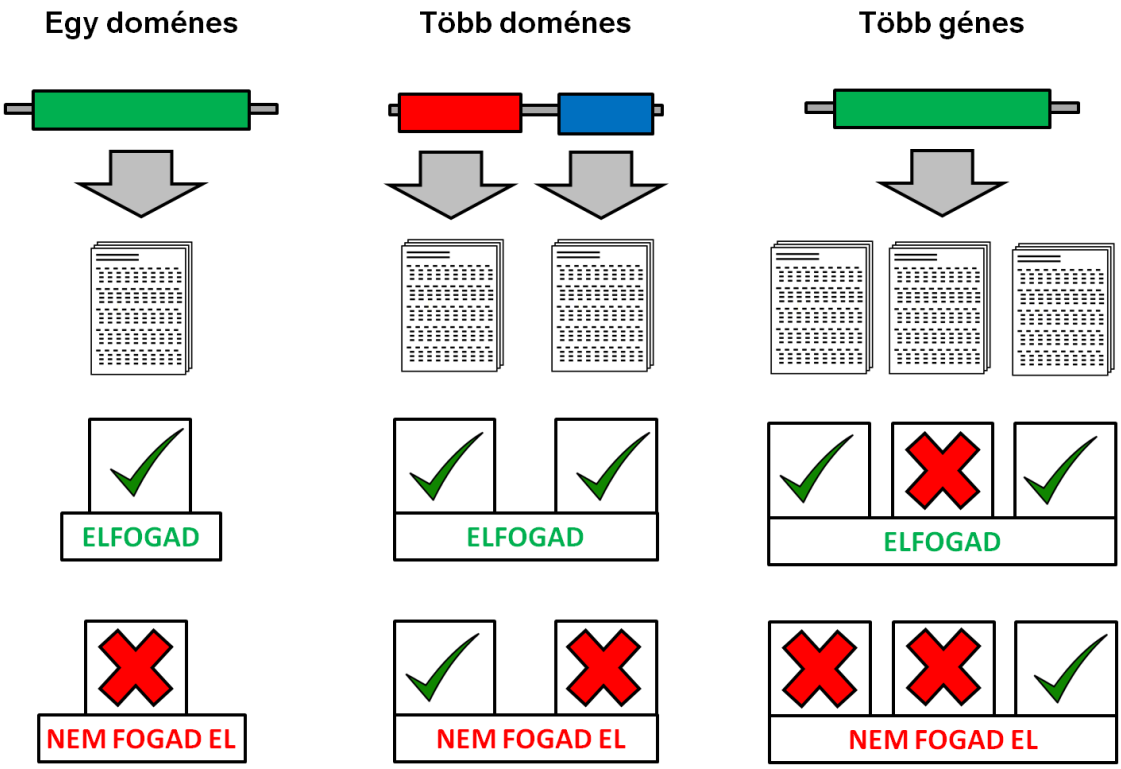
5.2.3. A keresés találatainak szűrése

A következő lépés a találatok feldolgozása a géncsalád információs fájlja alapján. Ez a feldolgozás két fontos feladatot lát el; kiszűri a túl valószínűtlen találatokat, illetve a talált domének alapján megállapítja, hogy az adott fehérjecsaládról van-e szó, vagy sem.

A HMM keresés eredményei esetén több lehetőségünk is van a valószínűtlen találatok kiszűrésére. Az egyik általam használt diszkriminátor a találat hossza. Minden gén esetén van egy jellemző hossz, amelytől a fehérjecsalád tagjai csak kis mértékben térnek el. Mivel a HMM profil készítéséhez szükségünk volt ismert családtagokra, így ezek hosszából már könnyen prediktálhatunk egy határt, amely fölötti hossz esetén már alacsony annak a valószínűsége, hogy a géncsalád egy tagjával van dolgunk. Ez a határ például lehet az ismert génhosszak maximumának bizonyos százalékkal történő növelése. A hossz vizsgálata az adatbázis rekordok hibáinak kiszűrésénél is hasznos, ugyanis előfordulhat, hogy az adatbázisban egy génként szereplő szekvencia valójában több gént tartalmaz. Ekkor a HMM keresés a részleges, de nagy egyezés miatt a géncsaládba tartozást fogja megállapítani, viszont mivel a „génünk” kétszer olyan hosszú, mint a többi, ezért itt kifog szelektálódni. Ez ugyan egy valószínűleg helyes gén elvesztését jelenti, amely csak rosszul szerepel az adatbázisban, de a végső eredmény validálásának megkönnyítése érdekében az adatbázis anomáliák esetén inkább a teljes kihagyást választottam.

A domén szintű vizsgálatra azért van szükség, mert előfordulhat, hogy a keresett fehérjecsaládunk egyik doménje egy rendkívül gyakori domén. Ebben az esetben a valószínűség vizsgálat során a szekvencia ezt a részét sokkal kisebb súllyal kell figyelembe venni, mert különben ez a DNS szakasz a találatok valószínűségi értékeinek eloszlástartományát szűkíti. A vizsgálat során 3 típusú géncsalád felírást használtam (5.9. ábra)

DOI:10.15774/PPKE.ITK.2014.006



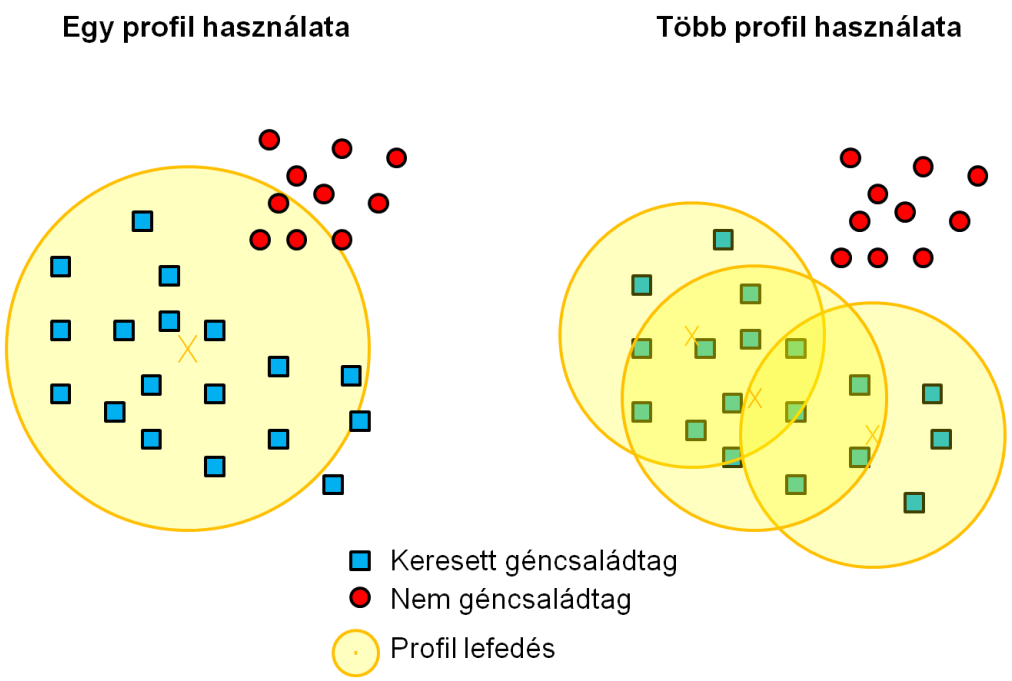
5.9. ábra A HMM profil lehetséges típusai

a) Egy doménes: Tulajdonképpen az egész szekvenciát írjuk fel egy HMM profillal. Ha a keresés szignifikáns eredményt mutat, a gént a fehérjecsalád részének tekinthetjük. Ilyet használtam a *rsaM* és *rsaL* gének esetén.

b) Több doménes: A szekvencia több jellemző domént tartalmaz, így mindegyiket külön HMM profillal írjuk fel. Csak akkor fogadjuk el a fehérjecsaládba tartozást, ha az adott szekvencia mindegyik domén esetén szignifikáns találatot mutat. Ilyen felírást alkalmaztam a *luxR* gén esetén a következő doménekkel: egy *autoinducer* kötő domén, és egy GerE nevű regulátor.

c) Több génes: Ezt abban az esetben kell használni, ha az adott géncsalád tagjai változatosak: nincsenek jellemző, diszkrinatív domének, ráadásul egy másik géncsalád nagyon hasonló. Ekkor az egész szekvenciát leíró HMM profil a biztos találatokra ugyan jól működne, viszont a kevésbé jó találatok esetén már rengeteg hibás találat is lehetséges. Ezért nem egy, hanem több, a teljes szekvenciát lefedő HMM profilt készítünk, és akkor fogadjuk el a találatot, ha ezek a profilok megadott százalékban szignifikáns hasonlóságot mutatnak. Ezt szemlélteti az 5.10. ábra. Ezt a módszert használtam a *luxI* gén esetén.

DOI:10.15774/PPKE.ITK.2014.006

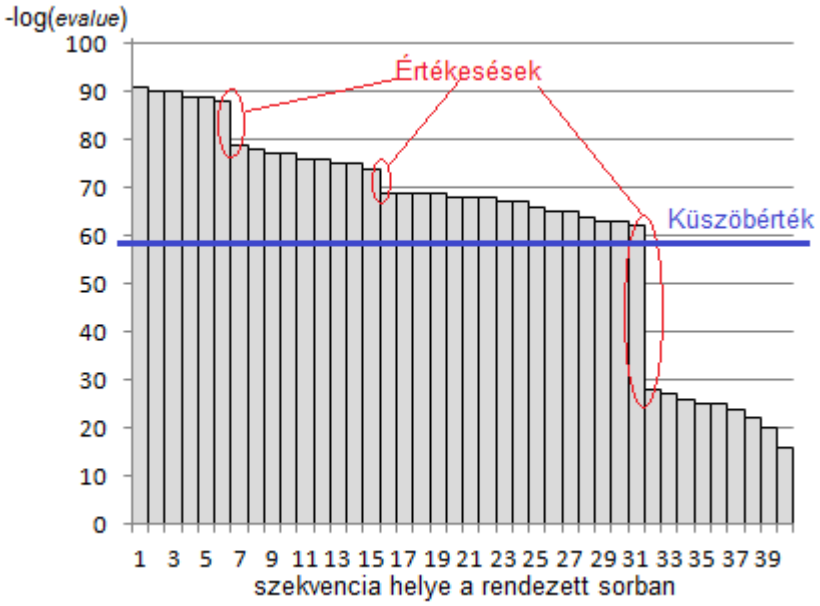


5.10. ábra Egy HMM profil lefedésének szemléltetése

Ha géneket egy két dimenziós térben ábrázoljuk, akkor egy HMM profil szignifikáns találatait tekinthetjük egy meghatározott sugarú körnek. Bizonyos esetekben előfordul, hogy egy profil használata esetén nem fehérjecsaládbeli gének magasabb pontszámú eredményt érnek el a keresés során, mint a fehérje család egyes tagjai. Ez a probléma több, specifikusabb profil alkalmazásával sok esetben megszüntethető.

A domén találatok szűrésének legfontosabb szempontja a vágási küszöbérték megválasztása, azaz meghatározzuk azt a találati valószínűséget, aminél jobb eredményeket már szignifikánsnak tekintünk. Ennek meghatározásához szükségünk van a találatok analízisére. (Amennyiben a géncsalád ismert tagjainak száma elég nagy, ez a művelet előre elvégezhető.) Először nézzük a HMM keresés adott doménjére vonatkozó találatok listáját, majd vesszük a találatok *e-vaule* értékét, növekvő sorrendbe rendezzük őket, és grafikonon ábrázoljuk a negatív logaritmusukat. (5.11. ábra) Ezen az ábrán jelentős értékeséseket keresünk, majd ezek közül választunk egyet, és az esés intervallumán belül választunk egy küszöbértéket (praktikusan a maximum értékéhez közeli értéket érdemes választani). Ha nincsen jelentős értékesés, akkor a HMM profilunk nem használható megbízható keresésre, mivel a találatok középső részében a fehérjecsalád tagjai és a nem tagok keverten helyezkednek el, így bárhogy is választjuk meg a küszöbértéket, vagy a fals pozitív vagy a fals negatív találatok száma lesz magas. Ebben az esetben vagy másik profilt kell készítenünk, vagy meg kell vizsgálnunk a profil felbontásának lehetőségeit. Minél alacsonyabb küszöbértéket választunk, annál valószínűbb a találatok helyessége, viszont annál több jó találatot zárunk ki a további vizsgálatból. Én a munkám során inkább magasabb küszöbértéket választottam, és az eredményt különböző validálási eszközökkel ellenőriztem.

DOI:10.15774/PPKE.ITK.2014.006



5.11. ábra Egy HMM keresés szignifikancia vizsgálatának egy példája

Az ábrán pirossal bekarikázva találhatóak a nagy meredekségű csökkenések. A kék vonal jelzi a választott szignifikancia szintet, ami a példában a grafikon 58-es szintje, ami 10^{-58}-os értéknek felel meg.

5.2.4. Adatok gyűjtése, találatok ellenőrzése

Miután csak a szignifikáns gének maradtak a listában, a program újfent csatlakozik az NCBI adatbázisához, hogy az elrendeződés vizsgálatához szükséges adatokat összegyűjtse: nevezetesen az eddig ismert génazonosító mellett a gén sorszámát, pontos helyét a genomban, melyik szálon található a gén, milyen COG csoportba van sorolva, és milyen fehérjét termel. Az utolsó kettő információ a találatok ellenőrzését könnyítik meg. Például ha tudjuk, hogy a keresett fehérjecsaládunk melyik COG csoportba tartozik, és a talált génnek COG csoportja ezzel megegyezik, az megerősíti a találat helyességét. Sajnos sok esetben hiányzik ez az érték, így nem lehet csak erre alapozni az ellenőrzést. A fehérje termék leírása is egy jó validálási pont, de sajnos a 2.6 fejezetben megfogalmazott nehézségek miatt ez sem lehet abszolút validálási tulajdonság.

Az adatokat egy másik módszer segítségével is ellenőriztem. Munkám kezdetén a HMM keresés alternatívájaként felmerült a BLAST algoritmus alkalmazása is (blastp). El is készült egy kezdetleges adatbázis a kezdetben ismert *luxR* és *luxI* génekből, amihez a keresésben résztvevő fehérjeszekvenciákat hasonlíthatjuk. Ez a keresés kezdetben reményteljes eredményeket hozott, de ezeknek az adatoknak a feldolgozását nagyban nehezítette, hogy a találatok nem pontosan illeszkedtek a génekre. Ez nem jelent problémát, ha a találatok végpontjai csak kis mértékben térnek a gének végpontjaitól, mert akkor könnyen megfeleltethetjük őket egymásnak. De több esetben a génnek csak kis részét, vagy egyszerre

DOI:10.15774/PPKE.ITK.2014.006

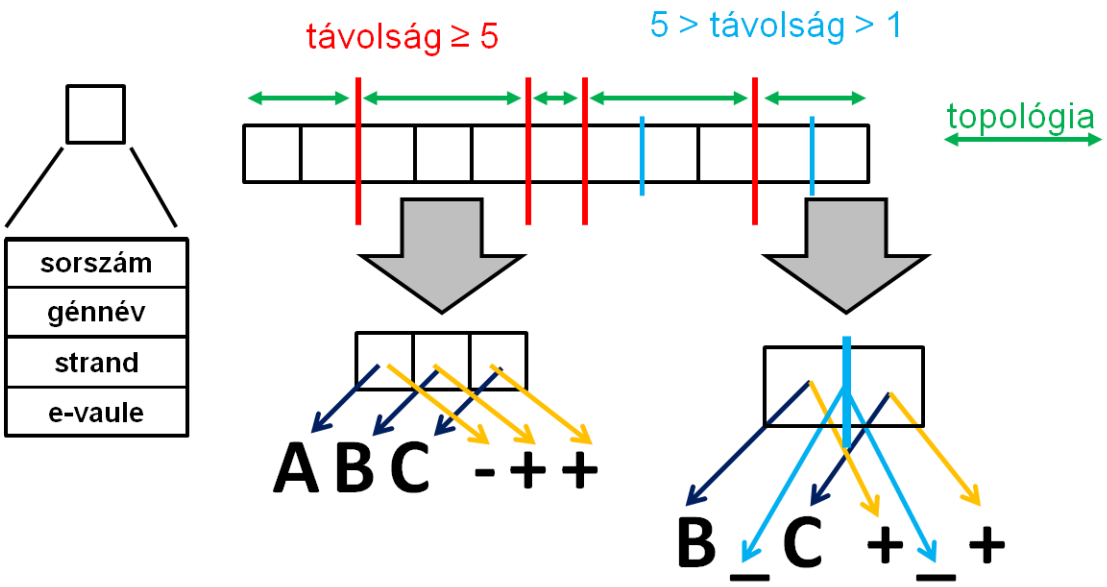
több gént is lefedett a találat, így ezeket a találatokat nem lehet további vizsgálat nélkül elemezni. Ezért döntöttem inkább a HMM alapú keresés mellett. A BLAST keresés eredményét azonban a későbbiekben alkalmazni tudtam az eredmények validálásban (bár az adatbázisát nem javítottam tovább): ha egy BLAST eredmény nagyrészt lefedi a génjelöltünket, azzal erősíti annak helyességét.

5.2.5. Topológiák keresése

Miután ellenőriztük a kibővített információjú találatainkat, nincs is más dolgunk, mint megvizsgálni, hogy az adott fehérjecsaládok génjei milyen helyzetben vannak egymáshoz viszonyítva. Mindegyik gén kapott egy rövidített jelzést (egy betűt), hogy könnyebben fel lehessen írni a talált topológiákat, azaz az elhelyezkedéseket. (*quorum sensing* gének esetén ez a következő volt: *luxR* – R, *luxI* – I, *rsaM* – M, *rsaL* – L).

A topológiák felismerésének első lépése, hogy a megtalált géncsaládtagokat baktérium kromoszóma szerint külön csoportosítjuk, majd ezután a genomban való szereplés sorszáma alapján sorba rendezzük. Mivel akkor tekintünk egy géncsoportosulást topológiának, ha nem tartalmaz 5 vagy annál nagyobb génhézagot, ezért a génsorunkat szétbontjuk minden 5-nél nem kisebb szakadásnál. Az így kapott részek már maguk a topológiák, amiket külön kezelünk. A topológia tagjainak adatai közül a gén rövidített neve és *strand* azonosítója segítségével felírjuk a topológiai elrendeződését. A sorozatból hiányzó gének helyét kihagyjuk, mivel később ezeket újabb keresések vagy más forrásokból pótolni tudjuk, illetve topológiák összehasonlításánál fontos különbség lehet a közbeékelődő gének száma. (5.12. ábra) Mivel a későbbiekben szükségünk lehet a topológia, úgymond „jóságára”, ezért az őt alkotó gének *e-vaule* értékeiből mértani átlagot számolva a topológiának is adunk egy valószínűségi értéket, amely a későbbiekben jelzi nekünk, hogy mennyire megbízható az adat. A topológiákat ezután egy szöveges fájlba tárolja el a program, amely a topológia típusa és valószínűségi értéke mellett tartalmazza a baktérium kromoszómájának azonosítóját, amiben a topológiát találtuk, és a topológia pontos helyzetét, kezdő és vég génsorszámmal megadva. Ez a későbbi vizsgálatoknál megkönnyíti az egy baktérium kromoszómán található topológiák vizsgálatát.

DOI:10.15774/PPKE.ITK.2014.006



5.12. ábra Az adott bakteriális kromoszómán talált topológiák kinyerése.

Az ábrán a sorba rendezett gének távolságát három csoportba foglaltam, és különböző színű függőleges vonalakkal jelöltem. Legalább 5 hosszú génhézagot piros, az 5-nél rövidebb génhézagot kék színnel jelöltem. Ha semmilyen jelölés sincs két gén között, akkor azok szomszédosak, azaz nincs közöttük semmilyen másik gén.

5.2.6. A keresésben részt vett HMM profilok felépítése

Mind négy fehérjecsalád esetén kezdetben egy kis létszámú szekvencia csoport felhasználásával készítettem el a HMM profilt. *LuxR* és *LuxI* fehérjék esetén ez körülbelül 20, míg *RsaM* és *RsaL* fehérjék esetén ez 5 illetve 6 fehérjeszekvenciát jelentett. A regulátor fehérjék esetén ezek a profilokkal végzett keresések megfelelő eredményt adtak, így ezeknek a fehérjecsaládoknak a keresése egy doménes maradt. (5.9. ábra) A *LuxR* fehérjék keresése során kiderült, hogy az *autoinducer* kötő doménje nagyon gyakori más fehérjék esetén is, így a találatok pontosabb szűrése érdekében itt áttértem a több doménes HMM keresésre. (5.9. ábra) A domének szekvenciáihoz az eredeti 19 fehérje doménjeinek szekvenciáit használtam. A *LuxI* fehérje esetén a több génes keresés hozta a legjobb eredményt, így a munkám során azt használtam; 4 különböző HMM profilt készítettem, és ha legalább 3 eredmény elérte a meghatározott küszöbértéket, akkor fogadtam el az eredményt. A profilok elkészítéséhez szükséges szekvencia csoportok az első, szigorú szűrési beállításokkal futó, manuálisan ellenőrzött keresés különböző topológiai csoportjai adták: R1, R2, R3 és L1+M1. A szekvenciák pontos listáját a X. melléklet tartalmazza.

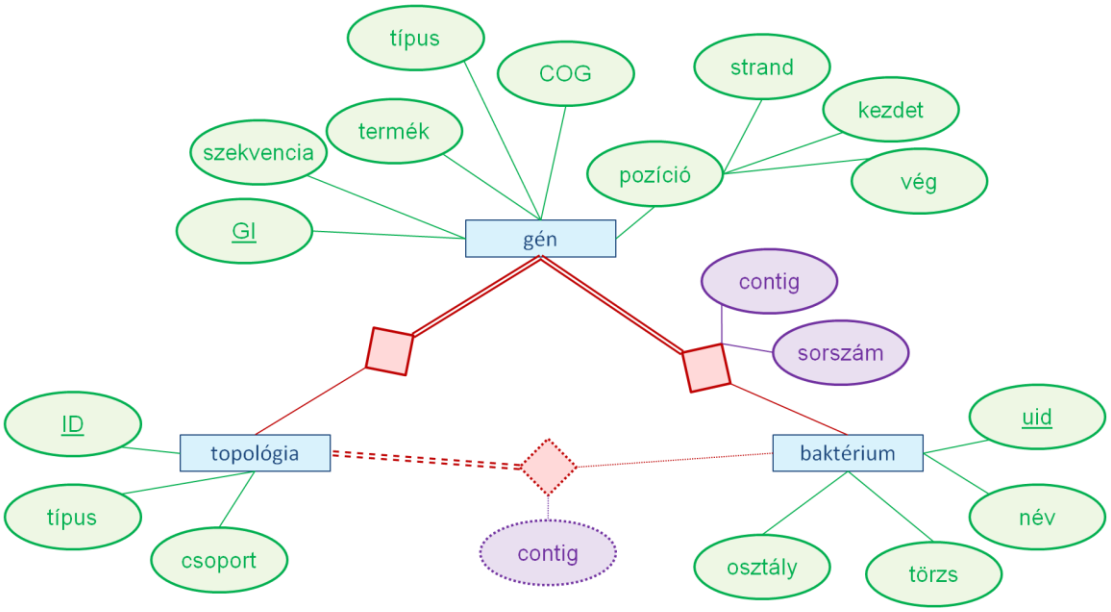
DOI:10.15774/PPKE.ITK.2014.006

5.3. Adatok tárolása

Az automata algoritmus lefutása után az eredményeket egy egyszerű szöveges fájlban kapjuk meg. Amennyiben kis keresési térben futattuk a programot, ez megfelelő volt a későbbi vizsgálatok forrásának. A teljes bakteriális genom adatbázison való futtatás azonban sokkal nagyobb méretű fájlt eredményezett, ami gyakorlatilag lehetetlenné tette, hogy az elemzés során felmerült kérdésekre kézi elemzéssel gyors választ kaphassunk. Az analízishez szükséges adatok gyors kinyerésének érdekébe az eredményeket egy relációs adatbázisban tároltam. Így lekérdezések írásával egyszerűen és gyorsan fértem hozzá azokhoz az információkhoz, ami az éppen felmerült kérdés megválaszolásához kellett. Például bizonyos tulajdonságú gének szekvenciáinak kigyűjtése vagy gének átfedésének vizsgálata.

5.3.1. A relációs adatbázis felépítése

Az adatbázisom elsődleges célja a benne tárolt információk gyors és egyszerű szűrésének lehetősége. Az adatok nagy száma miatt azonban figyelnem kellett arra is, hogy lehetőleg ne tároljak semmilyen redundáns információt. Ez azért is fontos, mert habár a rendszer egy kisebb, 4 elemből álló fehérjecsalád-csoport tárolására lett kitalálva, a későbbiekben előfordulhat ennél nagyobb elemszámú csoport is, ami esetén a redundancia komoly problémát okozhat, és teljes átrendezést tett volna szükségessé. A rendszerben szigorú megszorításokat lehet meghatározni, mint például minden génhez kötelező rendelni olyan baktérium azonosítót, ami szerepel az adatbázisban.



5.13. ábra Az algoritmus eredményeit tartalmazó adatbázis entitás-reláció diagramja

DOI:10.15774/PPKE.ITK.2014.006

Mivel elsősorban az adatok hordozzák az információ értéket az adatbázisban, és nem a kapcsolatok, ezért relatíve kevés entitással megoldható a feladat; 3 fő adatcsoport van, amelyek egy-sok kapcsolattal kapcsolódnak egymáshoz (5.13. ábra). Mindegyik entitás más biológiai szintű információt hordoz: közvetlenül a génnel és annak szekvenciájával kapcsolatban álló információ, a baktérium és annak taxonómiája és maguk a topológiák. A kapcsolatok mind 3 esetben a tartalmazást fejezik ki: a gén melyik topológiában van, a gén vagy topológiai melyik baktériumban található meg. Bár a topológia és baktérium közötti kapcsolat nem szükséges, mert a génadatok segítségével kinyerhető, nagyban megkönnyítette az adatok keresését. A baktérium entitás kapcsolatai esetén szükség van egy extra információra, a *contig* azonosítójára, amely megmondja, a baktérium melyik kromoszómáján vagy plazmidján található az adott gén vagy topológia. Gének esetén a génsorrendbeli helyzet is eltárolásra kerül, így az azonos kromoszómán lévő topológiák egymáshoz viszonyított helyzete is könnyen vizsgálhatóvá válik.

5.1. táblázat A relációs adatbázis entitásai és azok attribútumai.

gén entitás		
GI		A gén GI azonosító száma
szekvencia		A gén aminosav szekvenciája
termék		A gén által termelt fehérje neve (ha ismert)
típus		A kereső algoritmus által meghatározott fehérjecsalád neve (pl: *rsaM*)
COG		COG klaszter azonosítója
pozíció	strand	A gén melyik szálon helyezkedik el
	kezdet	A gén kezdő pozíciója a genomszekvencián (bázispár)
	vég	A gén vég pozíciója a genomszekvencián (bázispár)
topológia entitás		
ID		Egyedi szám, ami segít hivatkozni a topológiára
típus		A topológia típusa
csoport		A topológia nagyobb csoportja, például RI, RXI
baktérium entitás		
uid		Baktériumazonosító (NCBI)
név		A baktérium neve
törzs		A baktérium melyik taxonómiai törzsbe tartozik
osztály		A baktérium melyik taxonómiai osztályba tartozik

DOI:10.15774/PPKE.ITK.2014.006

5.4. Az eredmény megjelenítése

Miután az eredményeket sikeresen megkaptam és eltároltam egy relációs adatbázisban, már csak egy vizualizációs formát kellett választanom, hogy az adatok könnyen és gyorsan elérhetőek legyenek az informatikában kevésbé jártas szakemberek számára is. Mivel a cél a folyamatosan frissülő eredmények elérése, ezért a legmegfelelőbb megjelenítési formának egy honlapot tartottam.

5.4.1. A honlap keretrendszere

Először el kellett határoznom, hogy milyen eszközök segítségével készítem el a honlapot. Mivel mindenképp dinamikus weboldalra volt szükségem, így egy szerver oldali programozási nyelvet kellett választanom. A választásom a php nyelvre esett, mivel mind szintaktikája, mind szemantikája hasonlít az általam ismert Python és C++ nyelvekhez, és támogatja a relációs adatbázis közvetlen elérését is. A megjelenítés módjának kiválasztását nagyban könnyítette, hogy egyszerű szöveges információkat kellett megjelenítenem, így nem volt szükségem komoly grafikai megjelenítést támogató eszközökre (például: *Javascript, Flash*), hanem sima HTML (*HyperText Markup Language*) oldalakkal megvalósíthattam a kliens oldali vizualizációt. Mivel az oldalak elrendezése és színkészlete a felhasználók igényeihez kell hogy igazodjon, az egyszerű módosítás lehetősége miatt a stílust külön CSS (*Cascading Style Sheets*) fájlban tároltam.

5.4.2. A honlap felépítése

A weboldalon a megtekinteni kívánt adathalmaz kiválasztása után a főtáblázatot láthatjuk. (VI. melléklet) Az oldal tetején az aktuális adathalmaz nevét láthatjuk, míg alatta táblázatos formában az eredmények összesítését baktérium fajokra lebontva. Az első oszlopok a különböző topológiák adott fajban talált darabszámát mutatják, míg az utolsó pár oszlop az adott gének összdarabszámát tartalmazza.(5.14. ábra) A táblázat fejlécsora segítségével bármelyik oszlop szerint rendezni tudjuk az adathalmazt növekvő vagy csökkenő sorrendbe. A baktérium nevére kattintva tovább léphetünk az adott faj oldalára.

Change Data — **Datatable: QS2013** — Change Data

uid	name	proteo	sR	SI	sL	sM	R1	R3	R2	R4	L1	M1	M2	XX	LuxI	LuxR	RsaL	RsaM
58377	Rhizobium etli CFN 42	alpha	6	1			1			1				1	3	10		
58249	Agrobacterium vitis S4	alpha	8	3			1								4	9		
176372	Sinorhizobium meliloti Rm41	alpha	4	1			1							2	3	9		
58081	Burkholderia thailandensis E264	beta	3				1					1		2	3	8		2

5.14. ábra A főtábla címe és fejléce quorum sensing gének esetén.

DOI:10.15774/PPKE.ITK.2014.006

A baktérium oldala egy vagy több kisebb táblázatot tartalmaz, attól függően, hogy hány különböző *contig*jában találtunk *quorum sensing* gént. (VII. melléklet) Mindegyik ilyen táblázat az adott *contig*ban talált topológiákat sorolja fel, pontos helyük, típusuk és a valószínűségük feltüntetésével. Itt lehetőségünk van kiválasztani egy topológiát és tovább léphetünk a topológia oldalára. Ha nem csupán egy topológia génjeinek elhelyezkedésére vagyunk kíváncsiak, hanem a topológiák egymáshoz viszonyított helyzete is érdekel minket, akkor a *contig* nevére kattintva a *contig* oldalára léphetünk.

A topológiák oldalán az adott topológia génjeink adatait láthatjuk táblázatos formában, a kromoszómán való megjelenésük sorrendjében (VIII. Melléklet). Ha topológián belül egy hézag található, azaz a talált gének között egy olyan gén szerepel, ami nem volt tagja a keresés géncsaládjainak, akkor ez a gén egy vékony, adatokat nem tartalmazó sorral van jelölve, így felhívva a figyelmet a létezésére. A korábbi verziók esetén a felhasználóknak gondot okozott, hogy csak az első oszlopban szereplő sorszám utalt a hézagokra, így ezt a hangsúlyosabb jelölést kezdtem használni. Az oldal a génről megjeleníti a kereső algoritmus által kigyűjtött adatokat (pontos pozíció, GI és COG azonosító, termelt fehérje, a találat valószínűsége). A táblázat utolsó oszlopa egy linket tartalmaz, mely segítségével megnézhetjük az adott gén aminosav szekvenciáját mindenféle extra tartalom nélkül, így könnyen vágólapra helyezhető az információ. A *contig*ot megjelenítő oldal felépítése pontosan megegyezik a topológiák oldalával, csak az adott kromoszóma összes topológiáját megjeleníti egy táblázatban, vastag vonallal elválasztva őket egymástól. Az első oszlopban szereplő sorszámok segítségével, így lehetőségünk van a topológiák egymáshoz viszonyított helyzetét vizsgálni. Ez az ábrázolás bár nem olyan látványos, mint a kromoszóma térkép, viszont az utóbbival ellentétben sima szöveges honlapon is szépen megjeleníthető.

DOI:10.15774/PPKE.ITK.2014.006

6. Eredmények II. Baktériumok AHL QS génjei

6.1. AHL QS gének eloszlása a teljes bakteriális genomokban

Már korábban is ismert volt, hogy az azonos *quorum sensing* körhöz tartozó *luxI* és *luxR* gének vagy szomszédosak, vagy nagyon közel helyezkednek el egymáshoz a genomban. Ezenfelül a *luxR* géneknél előfordul olyan eset is, hogy nincs velük rokon N-AHL szintáz. Ezeket a géneket szóló [56] vagy árva [57] géneknek nevezik. A baktérium ezeknek a receptoroknak a segítségével figyeli a környezetében lévő más baktériumokat. Egy 2007-ben készült munka a *luxR* és *luxI* gének jelenlétét 512 genomban tárta fel, mely baktériumok mindegyike a proteobaktériumok törzsébe tartozik. [58] Emellett Goryachev készített egy összefoglalást a tandem és konvergens elrendeződésekről, melyeket A és B típusúnak nevezett el. [59, 60]

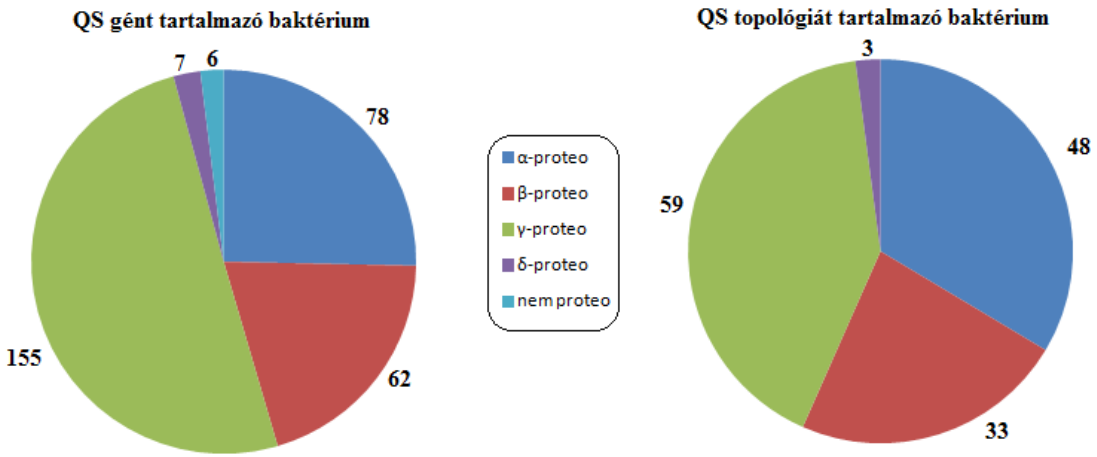
Az N-AHL alapú *quorum sensing* gének topológiai elrendeződésének elemzése a Pseudomonasok rendjébe tartozó baktériumok vizsgálatával indult, [13] melynek keresési terét kiterjesztettem az összes elérhető teljes baktérium genomra. (A bakteriális genomok forrása az NCBI GenBank adatbázisa volt.) A folyamat során 1403 genomot vizsgáltam meg standard bioinformatikai eljárásokkal, és eredményül 308 genomot találtam, amely tartalmazott *quorum sensing* gén, ebből 143 tartalmazott *luxR* és *luxI* gént is. Mindegyik baktérium a proteobaktériumok törzsébe tartozott. A baktérium osztályok eloszlását a 6.1. ábra és a 6.2. ábra mutatja be.

Mivel létezik mind a *luxI*, mind a *luxR* gének által kódolt fehérjéhez nagyon hasonló, más funkcióval rendelkező fehérje család, ezért a nem teljesen egyértelmű találati eredményeken manuális ellenőrzését a következő szempontok alapján végeztem el: hossz és szekvencia terjedelem. Az ellenőrzésekhez szigorú paramétereket állítottam be, hogy minél megbízhatóbb eredményhalmazt kapjak. Ezt a célt szolgálta a draft genomokból származó eredmények és az egyedül álló *quorum sensing* gének részletesebb analízise is. A nem annotált gének közül csak azok kerültek be az eredmények közé, amelyek egy ismert elrendeződés részei voltak. A 4,8 millió vizsgált bakteriális génben talált *quorum sensing* gének száma a következőképpen alakult: 674 *luxR* (33 nem annotált), 294 *luxI* (13 nem annotált), 44 *rsaL* (16 nem annotált) és 37 *rsaM* (egyik sem annotált).

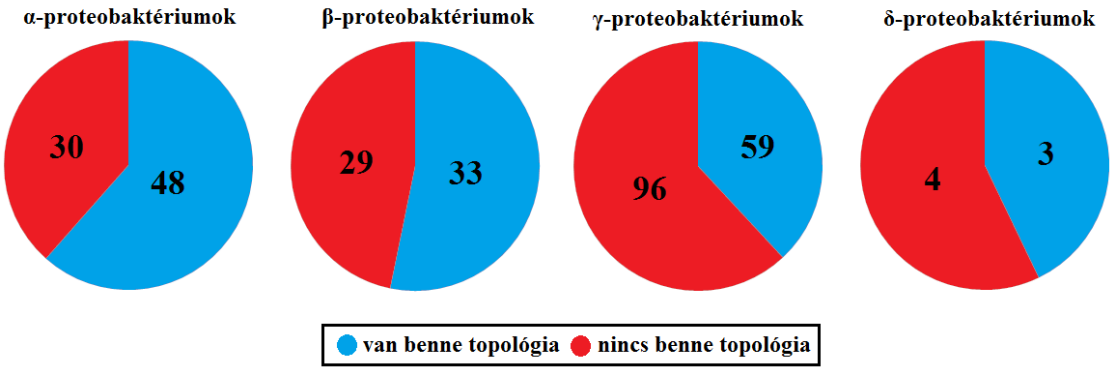
Felvetődik a kérdés, hogy a talált esetek tükrözik-e a *quorum sensing* gének természetben való megjelenésének frekvenciáját. Úgy gondolom, hogy ez nem teljesen van így. Ezt a következtetést több indokkal is alá tudom támasztani. Először is a vizsgálatunkat leszűkítettük azokra az esetekre, ahol a *luxR* és *luxI* gének egymás közelében helyezkednek el. Másodszor a keresés az ismert *LuxI* és *LuxR* fehérjékhez való hasonlóságon alapszik. Tehát

DOI:10.15774/PPKE.ITK.2014.006

kimaradtak azok a *luxR* gének, amelyek magányosan állnak vagy valószínű egy más típusú jeltermelést szabályoznak. Volt pár potenciális hasonlóság a proteobaktériumok törzsén kívül is, mint például a *Gloeothece PCC6909* nevű cianobaktérium esetén, amelynél korábban is felmerült, hogy *quorum sensing* rendszerrel rendelkezik. [61] Azonban úgy döntöttem, hogy a vizsgálatot a proteobaktériumok törzsére korlátozom, ahol a legtöbb jól megalapozott gén található. Harmadszor a vizsgálatot a teljes genomokon végeztem el, ami egy „elfogult" adathalmaz, és nem reprezentatív a természetben megtalálható összes baktériumra nézve. Ezekkel a megkötésekkel élve a proteobaktériumok 12%-ában találtam *quorum sensing* gént, ami összhangban van a proteobaktériumokban lévő AHL pozitív strainek frekvenciájával (6-12%). [58] Ennek az egyezőségnek a megerősítéséhez azonban további szigorú mintavételi eljárások szükségesek, és egy jóval nagyobb bakteriális genom adatbázis, ami jobban képes a természetben előforduló összes baktérium fajt reprezentálni.



6.1. ábra Quorum sensing gént tartalmazó baktériumok eloszlása



6.2. ábra QS gént tartalmazó proteobaktériumok topológia tartalmazása

DOI:10.15774/PPKE.ITK.2014.006

6.2. QS gének topológiai elrendeződése

A géntopológia egy általános kifejezés, mely a gének kromoszómán való elhelyezkedését jelenti, figyelembe véve a replikációs eredetet és más kromoszómális elemet. Jelen munkámban a topológiai elhelyezkedést vagy röviden topológiát a *quorum sensing* gének közeli szomszédságának elhelyezkedésére használom. Az elhelyezkedések illusztrálására egy PROSITE-szerű szintakszist dolgoztam ki. [62] A *luxR, luxI, rsaL* és *rsaM* géneket rendre R, I, L és M betűkkel rövidítem, egyéb gének esetén pedig az X-et használom. A génszimbólumok feletti nyíl pedig a transzkripció irányát jelöli. Ezzel a jelöléssel például az $\vec{R}\vec{I}$ egy szomszédos *luxR* és *luxI* génpárt jelöl, melyek azonos irányban íródnak át. Az átíródás iránya attól függ, hogy a gén a DNS melyik szálán helyezkedik el.

A talált mintákat két csoportra osztottam: egyszerű és összetett topológiák. Az egyszerűek egy *luxR* és *luxI* párt tartalmaznak, melyek vagy szomszédosak, vagy csak néhány gén található közöttük. Első közelítésben 0-3 közbenső génnel rendelkező topológiák tartoztak volna ebbe a csoportba, de a keresés után talált topológiáknál az esetek többségében csupán egy beékelődött gént találtam. Ha 1-nél több gén volt a *quorum sensing* gének között, akkor már legalább 4 vagy több.

Mivel a keresés elsődlegesen az egyszerű topológiák csoportjára irányult, ezt vizsgáltam részletesebben. Összetett topológia esetén a különbség csupán annyi, hogy a génpár között nagy mennyiségű egyéb gén található. Ezek a topológiák jellemzően az agrobaktériumok és rhizobium fajokban fordulnak elő, melyekről több összefoglaló cikk is fellelhető. [63, 64]

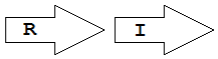
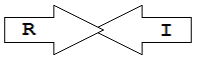
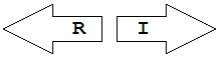
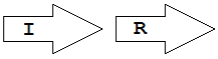
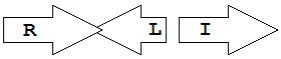
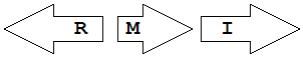
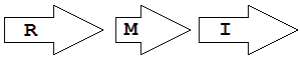
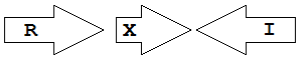
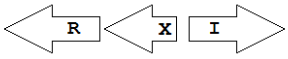
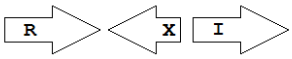
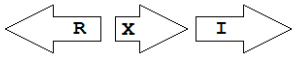
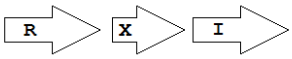
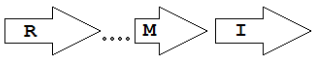
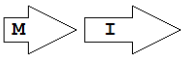
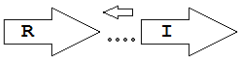
6.2.1. Azonosított topológiák

Az egyszerű topológiák között a két leggyakoribb elrendeződés az $\vec{R}\vec{I}$ (R1) és az $\vec{R}\overleftarrow{I}$ (R2) topológia, melyet Goryachev A és B típusúként nevezett. [59, 60] Viszont ezeken kívül még más topológiákat is találtam, így mind a négy, elméletben lehetséges két génből álló elrendeződés is megjelent az adatokon, bár az új típusúak csak sokkal kisebb számban.

A teljes baktériumok még alacsony száma nem engedi meg, hogy biztos következtetéseket vonjunk le a különböző elrendeződési minták megjelenéséről a különböző baktérium csoportokban és fajokban. Pár észrevétel azért tehető: az $\vec{R}\vec{I}$ topológia az α-proteobaktériumokban domináns, míg az $\vec{R}\overleftarrow{I}$ topológia a γ-proteobaktériumokban fordul elő leggyakrabban. Továbbá az $\vec{R}\overleftarrow{L}\vec{I}$ és $\overleftarrow{R}\vec{M}\vec{I}$ topológiák mind β, mind γ osztályok esetén előfordulnak, de α esetén nem. (6.1. táblázat)

DOI:10.15774/PPKE.ITK.2014.006

6.1. táblázat Topológiák eloszlása a proteobaktériumokban

ID	Minta	Topológia	Megjelenés a proteobaktériumokban				
			Összes	alfa α	béta β	gamma γ	delta δ
Rövid, konzervált topológiai minták							
R1	$\vec{R}\vec{I}$		*96*	71	14	11	0
R2	$\vec{R}\overleftarrow{I}$		*53*	2	2	46	3
R3	$\overleftarrow{R}\vec{I}$		*11*	1	3	7	0
R4	$\vec{I}\vec{R}$		*2*	2	0	0	0
L1	$\vec{R}\overleftarrow{L}\vec{I}$		*15*	0	7	8	0
M1	$\overleftarrow{R}\vec{M}\vec{I}$		*30*	0	20	10	0
M2	$\vec{R}\vec{M}\vec{I}$		*1*	0	1	0	0
X1	$\vec{R}\vec{X}\overleftarrow{I}$		*1*	0	0	1	0
X2	$\overleftarrow{R}\overleftarrow{X}\vec{I}$		*2*	2	0	0	0
X3	$\vec{R}\overleftarrow{X}\vec{I}$		*4*	0	2	2	0
X4	$\overleftarrow{R}\vec{X}\vec{I}$		*1*	1	0	0	0
X5	$\vec{R}\vec{X}\vec{I}$		*2*	1	1	0	0
Hosszú, szokatlan topológiai minták							
M3	$\vec{R}X(2-11)\vec{M}\vec{I}$		*6*	0	6	0	0
M'	$\vec{M}\vec{I}$		*2*	0	2	0	0
X6	$\vec{R}\overleftarrow{X}(7)\vec{I}$		*1*	1	0	0	0
X7	$\vec{I}X(>7)\vec{R}$		*5*	5	0	0	0

DOI:10.15774/PPKE.ITK.2014.006

6.2.2. A közbenső gének vizsgálata

A 48 darab egyszeresen beékelődő génnek több fajtája is van. 15 kódolja az *RsaL* és 31 kódolja a *RsaM* fehérjét, melyek ismert negatív szabályzói a *quorum sensing* mechanizmusnak. Mindkét esetben egy jellemző topológiát találtam: $\vec{R}\overleftarrow{LI}$ és $\overleftarrow{R}\vec{M}\vec{I}$. Az *RsaL* egy tetra helikális alosztálya a H-T-H fehérjének [65], - egy gyakori *quorum sensing* gátlóként ismert fehérje - amely dimerként kötődik a DNS-hez. Például az *RsaL* fehérje a *Pseudomonas aeruginosa* nevű baktériumban megakadályozza a *luxR* gén kifejeződését azáltal, hogy a *lux-box* közelében a DNS-hez kötődik.[14] Ezzel ellentétes a talált negatív szabályzó *RsaM* fehérje, melynek a pontos struktúrája nem ismert, és kizárólag *quorum sensing* körökben tűnt fel a vizsgálat során. [66] Habár legtöbb esetben az *RsaM* fehérje $\overleftarrow{R}\vec{M}\vec{I}$ topológiában jelenik meg, kis számban más elrendeződések is előfordultak (M', M2, M3). A maradék közbenső gén egy része vagy más típusú negatív regulátor szerepet tölt be (nem a transzláció gátlásával szabályozza a kommunikációs kört, hanem például a jelanyag lebontásával) vagy teljesen ismeretlen funkciójú. Az ismeretlen funkciójú gének között nem sikerült számot tevő hasonlóságot kimutatni sem a szekvenciákban, sem a topológiát alkotó gének orientációjában, így további csoportosításuk sem lehetséges. A 6.1. táblázat az sugallhatja, hogy az X1-X5 topológia típusok közül pár többször is előfordult, így hasonlóságot lehetne keresni az adott gének között, de minden esetben ugyanazon baktériumfaj különböző egyedeiben szerepelő, azonos *quorum sensing* körhöz tartozó gének voltak, így nem láttam szükségesnek a még részletesebb vizsgálatukat. A három tagból álló topológiák középső génjének eloszlását a 6.2. táblázat mutatja be.

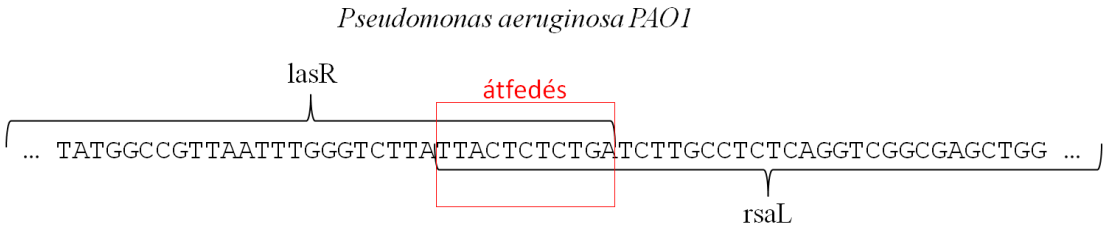
6.2. táblázat Közbenső gének a rövid, konzervált topológiákban

gén típus	talált gének száma	feltételezhető szerep	példa genom
RsaL	15	negatív regulátor	*P. aeruginosa PAO1*
RsaM	31	negatív regulátor	*B. pseudomellei K96243*
MupX	1	negatív regulátor	*P. flourescens NCIMB 10586*
integráz/transzpozáz	2	DNS mobilizáció	*B. vietnamiensis G4*
LuxR típusú regulátor	1	?	*Gluconacetobacter PAI5*
Ismeretlen	6	?	*B. mallei NCTC 10247*

DOI:10.15774/PPKE.ITK.2014.006

6.2.3. Gének közötti átfedések

A gének konzervált átfedése egy fontos leírója a topológiáknak. Az átfedések vizsgálatakor a gének kódoló régióit (CDS) vettem alapul. A munkám során két csoportnál észleltem ezt a jelenséget: $\vec{R}\vec{L}\vec{I}$ (L1) és $\vec{R}\vec{I}$ (R2). Az L1 típusú topológia esetén az átfedés az ellentétes irányban átíródó *luxR* és *rsaL* gének között figyelhető meg. Az átfedés mértéke változó; *Pseudomonas aeruginosa* esetén 10 bázispár, míg *Pseudomonas fuscovaginae* esetén 20 bázispár. Ezzel ellentétben, *Pseudomonas putida* esetén a *luxR* és *rsaL* gének habár közel vannak egymáshoz (4 bázispár), még sincs átfedés. [66] Az R2 típusú topológiák 2 és 79 bázispár közötti átfedést tartalmaznak az ellentétes irányban kifejtődő *luxR* és *luxI* gének között. Korábban felmerült, hogy az egyik ilyen gén akadályozza az átfedő pár másik génjének kifejtődését, ezzel különböző funkció vagy fenotípus aktiválásokat vagy elnyomásokat eredményez. [67] Ez a hatás vagy a második RNS polimeráz molekula sikertelen felismerése által történik, vagy a két mRNS hibridizációja által. Az átfedő gének nem szokatlanok a szorosan szabályozott bakteriális génkörökben [68], mint például megszorítás módosító rendszerekben [69], de a konvergensen kifejtődő, átfedő gének kevésbé elterjedtek az irodalomban.



6.3. ábra Génszekvenciák átfedése

A *Pseudomonas aeruginosa PAO1* baktérium L1 típusú topológia génjeinek kódoló régióinak 10 bázispár hosszú átfedése a 1 558 880 és 1 558 890 pozíció között.

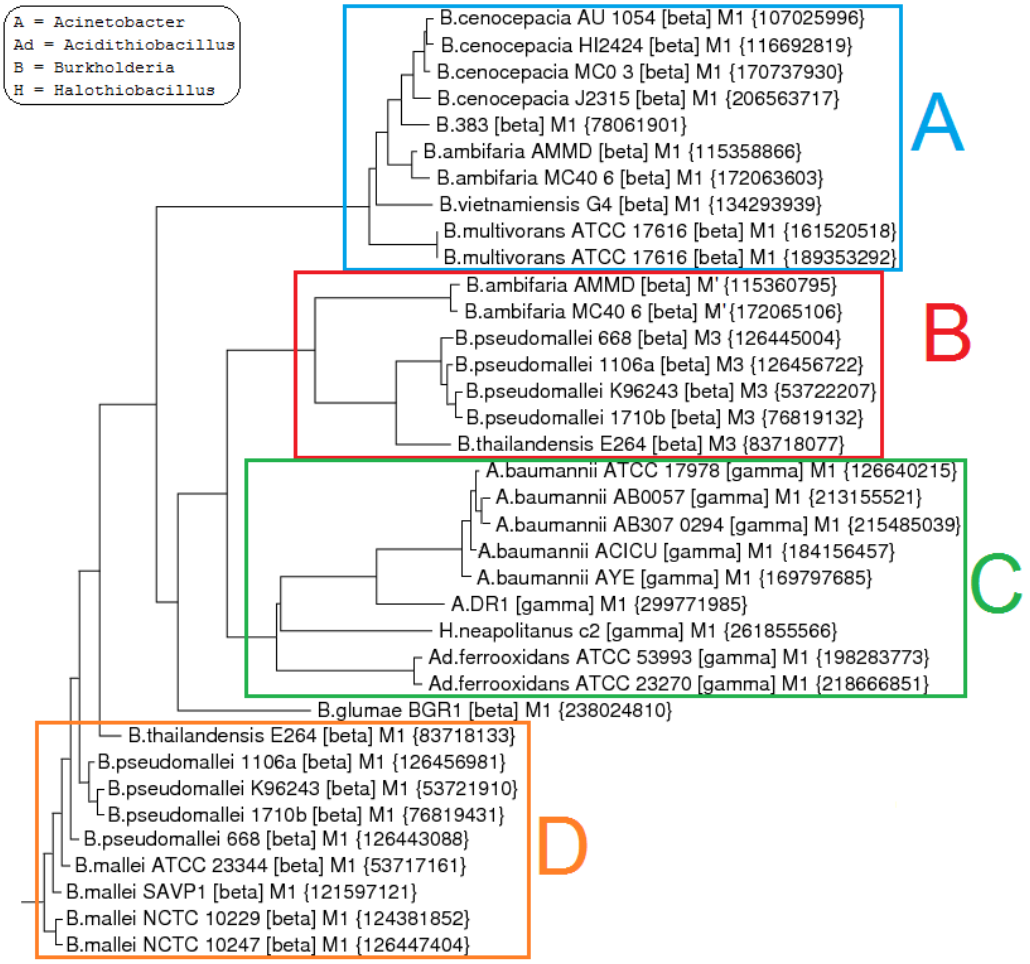
6.2.4. Hosszú topológiák

Az egyszerű topológiákkal ellentétben a hosszú, komplex topológiák halmaza sokkal nagyobb változatosságot mutat a talált esetek kisebb száma ellenére is. Az *agrobaktérium*okban és különböző *rhizobia* baktériumokban a *luxR* és *luxI* gének között nem csak egy, hanem több gén is található. Egy érdekes példa a *Burkholderia ambifaria* **MI** topológiája, ahol a két *quorum sensing* gén (M és I) tandem megjelenése mellett a vizsgált határon belül nem található annotált, vagy keresés által azonosított *luxR* homológ. Ez az eredmény azonban nem zárja ki annak a lehetőségét, hogy a *LuxR* fehérjecsalád egy általam nem vizsgált tagja található az adott helyen.

DOI:10.15774/PPKE.ITK.2014.006

6.3. A topológiai minták taxonómiai eloszlása.

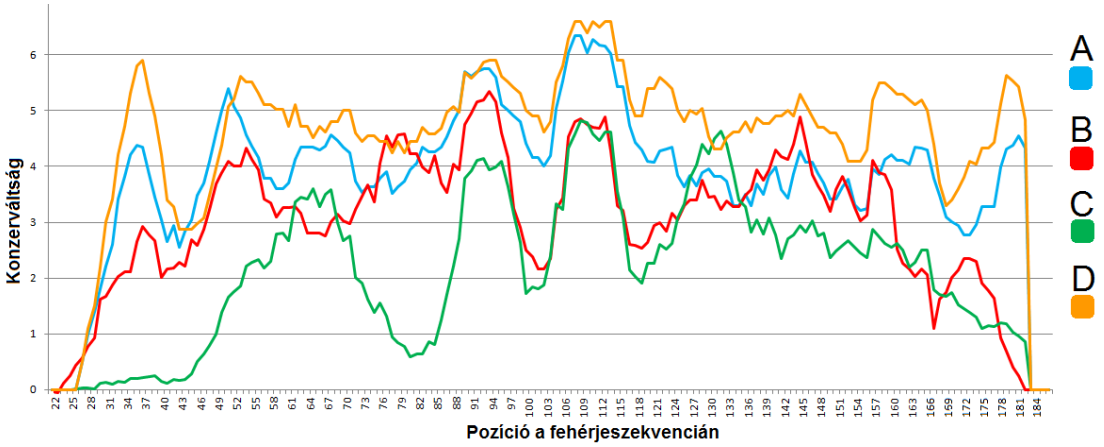
A 162 *LuxI* és *LuxR* fehérjeszekvenciáiból készült kladogramok elemzése azt mutatta ki, hogy a különböző topológia típusokban szereplő fehérjék tisztán megkülönböztethető csoportokra szeparálódnak. Ezen fatípusú gráfok helyett a méretre való tekintettel, egy könnyebben átlátható példát mutatok: az *RsaM* fehérjék kladogramját. (4.4. ábra) A filogenetikus fát megvizsgálva azt tapasztalhatjuk, hogy mind a topológia típusa, amiben az *rsaM* gén szerepel, mind a baktérium fajok rokonsága számított a fa által kapott csoportok kialakulásában. Az ábrán látszik a topológiák különválása a fában. Két csoport van; **M1** és **M3/M'**. Látható, hogy a különböző topológiákból származó gének nem keverednek egymással: az **M3/M**' típusú gének a piros (**B**-vel jelzett) csoportban találhatóak, míg az **M1** típusú gének a maradék háromban. Az ábrán az is látszik, hogy a baktérium osztályhoz való tartozás is befolyásolta - ugyancsak másodlagosan - a fában való elhelyezkedést. A γ-proteobaktériumok a **C** jelű, a β-proteobaktériumok az **A** és **D** jelű csoportban jelentek meg.



6.4. ábra Az rsaM gének klaszterеződése a kladogramjuk alapján

DOI:10.15774/PPKE.ITK.2014.006

Az *rsaM* gének klaszteresződésének megfigyelése után megvizsgáltam, hogy van-e az adott csoportokra jellemző konzerváltsági mintázat. Ehhez az 5.1.4 fejezetben említett konzerváltsági ábrát hívtam segítségül. Az *rsaM* gének többszörös illesztését 4 részre bontottam az adott csoportba való tartozás alapján. Azért nem a csoportokra külön-külön végeztem el az illesztés, mert akkor különböző hosszúságú illesztéseket kapnék, és azokat nem lehet összehasonlítani egymással. A közös illesztés viszont garantálja az azonos méretet. A négy darab, szétbontott illesztés egy közös grafikonon ábrázoltam. (6.5. ábra)



6.5. ábra Az rsaM génklaszterek konzerváltság ábráinak összehasonlítása

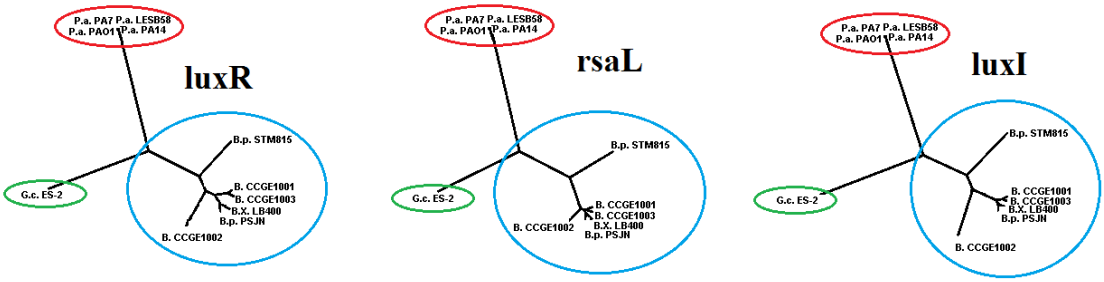
Az grafikonon ugyanazokat a színeket használtam a csoportok jelölésére, mint a kladogramon. Mivel az ábra elején konstans nulla értékek találhatóak, ezért az a részt levágtam. Ezért kezdődik a vízszintes tengely számozása 22-től.

Sajnos az ábra alapján nem lehet megalapozott állítások megfogalmazni. Ennek valószínű az az oka, hogy a csoportok viszonylag kevés (7-10) szekvenciát tartalmaznak, továbbá nem vizsgáltam, hogy az irányító fa további finomításával nem jelennek-e meg további alcsoportok. Ennek ellenére pár észrevétel tehető a grafikonnal kapcsolatban:

- Megfigyelhető, hogy mind négy csoport rendelkezik két, a környezetéhez képest viszonylag konzervált szakasszal a 94-es és a 112-es aminosav pozíció környékén. Ezen a két pozíció volt a legmagasabb a konzerváltság az összes *rsaM* gén felhasználásával készült konzerváltsági ábrán is. (5.8. ábra)
- Az is észrevehető, hogy az **A**, **B** és **D** csoportok rendelkeznek A 35-ös és 54-es aminosav pozíció környékén egy-egy konzervált régióval, addig a **C** csoportban ezek nem találhatóak meg. Helyette ebben a klaszterben a 67-es pozíció környékén található egy konzerváltabb szakasz.

DOI:10.15774/PPKE.ITK.2014.006

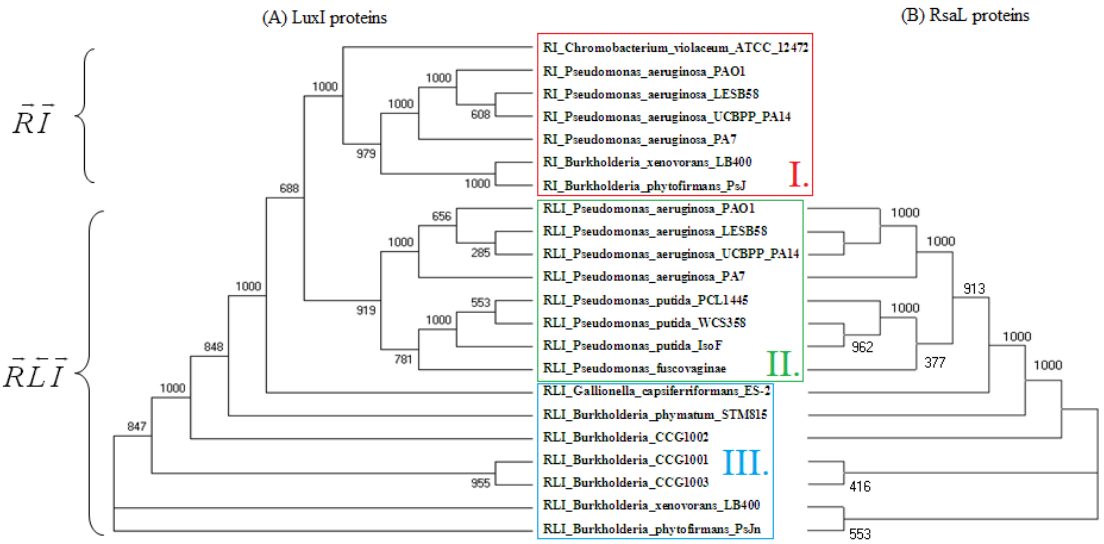
A kladogram elemzésének eredményei hasznosnak bizonyultak, így több különböző tulajdonság alapján szűkített fát rajzoltam. Például a *Pseudomonas* vagy a *Burkholderia* családhoz tartozó baktériumok *luxI*, *luxR*, *rsaM* illetve *rsaL* génjeinek felhasználásával. A fákon ugyancsak látszódtak a fentebb említett megállapítások, azaz az azonos topológiából származó gének elhatárolható csoportokat alkotnak. A csoportosulások egyértelműsége arra utalt, hogy a csoportok függetlenek a gén típusától. Ezt ellenőrizendő elkészítettem a topológiákat alkotó gének filogenetikai fáit külön-külön, majd megnéztem a fa által meghatározott csoportok elmeit. Azt tapasztaltam, hogy lényegében megegyeznek. A 6.6. ábra mutatja az topológiák összehasonlítását.



6.6. ábra L1 topológiák génjeinek ábrázolása gyökértelen fákon

A fák 11 darab L1 topológia felhasználásával készültek. A részt vevő baktériumok: *Pseudomonas aeruginosa PAO1*, *P. aeruginosa PA7*, *P. aeruginosa UCBPP-PA14*, *P. aeruginosa LESB58*, *Gallionella capsiferriformans ES-2*, *Burkholdéria xenovorans LB400*, *B. CCGE1001*, *B. CCGE1002*, *B. CCGE1003*, *B. phymatum STM815* és *B.phytofirmans PsJN*

Ezután megvizsgáltam, hogy a gének csoporton belüli elhelyezkedése mennyire egyezik meg a különböző gének esetén. A 6.7. ábra jól mutatja, hogy nem csak a csoportba tartozás, hanem a csoportokon belüli viszonyok és nagy hasonlóságot mutatnak.



6.7. ábra LuxI és RsaL fehérjék kladogramjainak összehasonlítása

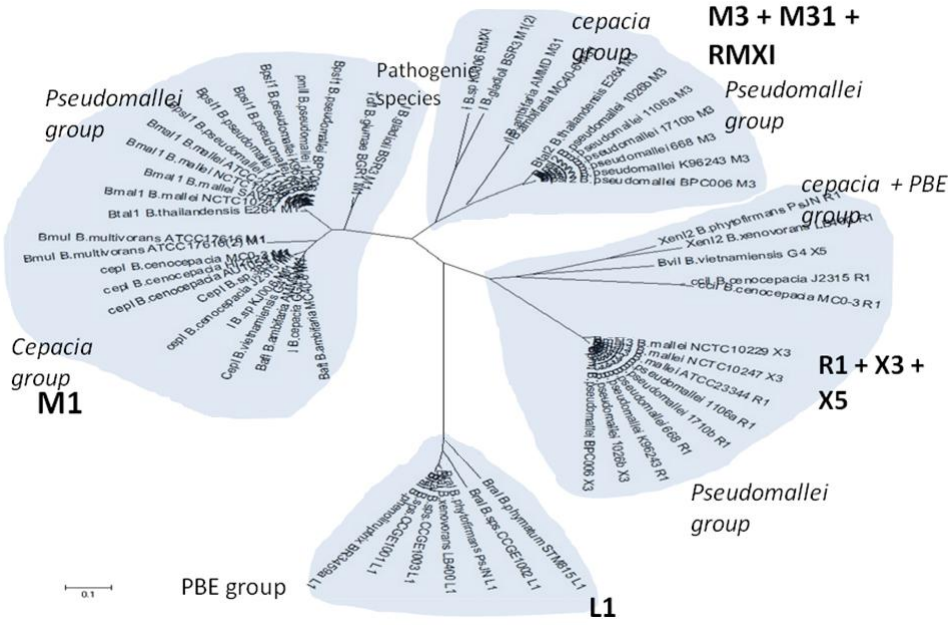
DOI:10.15774/PPKE.ITK.2014.006

6.3.1. Nagyobb elemszámú fák

A korábban mutatott két kladogram esetén a vizsgált szekvenciák száma viszonylag alacsony (<40). Ennek ok az ábrázolás részletessége. Már a 6.4. ábra csoportjainál is látható, hogy bizonyos ágak nagyon rövidek illetve össze is csúsznak. Ez nagyobb szekvencia szám esetén még gyakrabban fordul elő, ráadásul az ábra mérete is nő, és ez megnehezíti az ábra átlátását. Ugyan meg van arra a lehetőségünk, hogy a kladogramnak csak kisebb részét vizsgáljuk egyszerre, de ennél pontosabb képet kapunk, ha csak ebből a szűkebb szekvencia halmazból rajzolunk filogenetikus fát. Ezzel csökkentettem annak az esélyt, hogy a túlságosan nagy különbségek miatt a filogenetikus fa elveszítse informatív jellegét.

Mivel a keresés során talált *luxI* és *luxR* homológok száma több száz volt, így találnom kellett egy másik ábrázolási módot, amely segítségével nagyobb elemszámú szekvencia csoport is vizsgálható. A gyökérnélküli filogenetikus faábrázolás megfelelt ennek a célnak.

A teljes bakteriális genomokon való keresés után a topológiákat tartalmazó baktériumok közül sok tartozott a *Burkhodériák* rendjébe, és mivel ez egy jól ismert baktériumrend, így az ehhez tartozó fajokat külön is megvizsgáltam. A *burkholdéria* genomokban talált *LuxR* fehérjék csoportosulását a 6.8. ábra mutatja. A filogenetikus fán megjelenő csoportoknál két különböző dolgot vizsgáltam: az adott gén milyen topológiában vesz részt és a *burkholdéria* baktériumok melyik alcsoportjába tartozik. Az ábra szépen mutatja, hogy mindkét feltétel esetén a filogenetikus fa által mutatott csoportosulás megegyezik a főbb taxonómiai alcsoportokkal illetve csoportonként azonos típusú topológiában vesznek részt. Mivel minden *burkholdéria* baktériumban egyféle topológiából csak egy darab szerepelt, így a *luxR* homológok azonosítása a baktérium neve és a topológiai kódja alapján történt.



6.8. ábra A Burkholdéria baktériumokban található luxR homológok klaszterei

DOI:10.15774/PPKE.ITK.2014.006

6.4. A Pseudomonas törzs QS rendszerei

A *Pseudomonas* név a *Gram-negatív*, aerob *gammaproteobaktériumok* egy törzsét jelöli, melynek jelenleg 191 érvényesen leirt tagja van. A törzs tagjai a legjobban tanulmányozott baktériumok közé tartoznak, melyek anyagcseréje igen változatos, és nagyon sokféle környezetben tudnak élni.

A törzs tagjainak általános tulajdonságai, hogy Gram-*negatívak* (azaz csak vékony külső peptidoglikán sejtfallal rendelkeznek, amit lipopoliszaharid réteg fed), bot alakúak, egy vagy két poláris flagellájuk segítségével képesek helyüket változtatni, aerob körülmények között élnek és nem képeznek spórákat. Biokémiai tulajdonságuk a pozitív kataláz- és oxidázteszt. A *Pseudomonas* törzs tagjaira jellemző az úgy nevezett sziderofór molekulák termelése, ezek fémionok kötésére alkalmas struktúrák, amelyek segítségével a baktériumok a környezetből fel tudják venni a létfenntartásukhoz esszenciális fémionokat, pl. vasat. A legtipikusabb ilyen molekula a zöldessárgán fluoreszkáló pyoverdin, amelyet a törzs felismerésére is használnak.

A *Pseudomonas* törzs tagjaira jellemző a *biofilm*-képzés, és a *biofilm* struktúrákon belül az életkörülmények már nem felelnek meg az aerob leírásnak. A *biofilm* struktúrákon belül is már nagy a sejtsűrűség és az oxigén nem tud behatolni a struktúra teljes mélységébe. Jellemző tulajdonsága a *Pseudomas* törzs tagjainak az erős antibiotikum rezisztencia és az ugyancsak erős hajlam a horizontális géntranszferre. Ez a kettő együtt főleg azért aggályos, mert a *Pseudomonas* törzs tagjai szinte minden területen, azaz talajban, vizekben, állati és növényi gazdaszervezetekben megtalálhatók, így nagyon alkalmasak arra, hogy belőlük széleskörű rezisztenciával rendelkező új patogének alakuljanak ki.

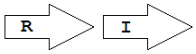
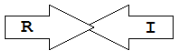
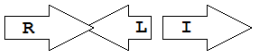
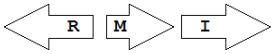
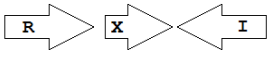
A törzs tagjait inkább életmódjuk szerint osztályozzák, tehát megkülönböztetnek állati és növényi patogéneket, növényekre jótékony hatást kiváltó törzseket és opportunista patogéneket, mint amilyenek a kórházak falán tenyésző *Pseudomonas aeruginosa* variánsok. Ez a faj többek között azért is ismert, mert a genetikus rendellenességből származó *cisztás fibrózis*ban szenvedő betegek tüdejét kolonizálja. Az életmód-szerinti csoportosítás mellett csak napjainkban kezdik elfogadni a DNS szekvencia alapján történő osztályozást, amely érdekes és ellentmondó eredményeket adott. Kiderült ugyanis, hogy a törzs nem egységes, hanem 7 jól definiált alcsoportra osztható, ezen felül marad néhány nem osztályozható faj. Ezeket a csoportokat azonban még nem használják általánosan a szakirodalomban.

A *Pseudomonas* törzs tagjainak kétharmada rendelkezik *quorum sensing* génekkel. Felmérésünk szerint ezek az előző fejezetekben ismert topológiák közül csak néhányat tartalmaznak (6.3. táblázat és 6.4. táblázat), a negatív regulációt biztosító gének közül inkább az *rsaL* gén a jellemzőbb, jelen dolgozat írása idején a nyilvános adatbázisokban csak egyetlen faj, a *Pseudomonas fuscovaginae* tartalmazza az *rsaM* gént. Megjegyzendő tulajdonság a

DOI:10.15774/PPKE.ITK.2014.006

gének átfedése, amelyet kétféle topológiában, az R2-ben és az L1-ben is megfigyelhetünk. Ilyen átfedések nem ismeretlenek a bakteriális genomokban, általában a szorosan együtt szabályzott géneknél fordul elő. Ez természetesen igaz a *quorum sensing* génekre is. Az R2 topológiai típusban az R és I gének átfedése 2 és 65 bázispár között van a *Pseudomonas syringae* csoportban, de a *Pseudomonas fluorescens* genomoknál nem találunk átfedést. Az L1 topológiai csoportban a *P. aeruginosa* fajoknál egységesen 10 bázispár az átfedés, a *Pseudomonas putidá*nál viszont nincs átfedés. Az R2 típusú topológia génjeiről már előzőleg leírták [67], hogy speciális regulációs mechanizmus szerint működnek, amely eltér a többi AHL *quorum sensing* rendszertől, amennyiben az R fehérje az AHL molekula távollétében köt a DNS-hez, és az AHL jelmolekula hatására válik le a DNS-ről. Ugyancsak Tsai és Winans írta le, hogy az átfedés önmagában negatív szabályzást jelenthet: vagy mert az átírást végző polimeráz molekulák akadályozzák egymás működését, vagy mert az átírás révén keletkező transzkriptek hatástalanítják egymást.

6.3. táblázat A Pseudomonas törzs tagjaiban szereplő topológiák

ID	topológia	Pseudomonas alosztályok*						Fajok
		PA	PC	PF	PP	PS	UG	
R1		4	0	0	0	0	0	P. aeruginosa LESB58 P. aeruginosa PA7 P. aeruginosa PAO1 P. aeruginosa UCBPP-PA14
R2		0	1	1	0	3	0	P. syringae 1448A P. syringae B728a P. syringae DC3000 P. chlororaphis PCL1391 P. flourescens 2-79
L1		4	0	0	3	0	1	P. aeruginosa LESB58 P. aeruginosa PA7 P. aeruginosa PAO1 P. aeruginosa UCBPP-PA14 P. putida WCS358 P. putida PCL1445 P. putida IsoF P. fuscovaginae UPB0736
M1		0	0	0	0	0	1	P. fuscovaginae UPB0736
X1		0	0	1	0	0	0	P. flourescens NCIMB 10586

*PA = Pseudomonas aeruginosa, PC = Pseudomonas chlororaphis, PF = Pseudomonas flourescens, PP = Pseudomonas putida, PS = Pseudomonas syringae, UG = nem csoportosított

DOI:10.15774/PPKE.ITK.2014.006

6.4. táblázat A Pseudomonas törzs tagjainak quorum sensing körei

Pseudomonas faj	***quorum sensing* kör**	**topológia**
P. aeruginosa LESB58	rhlR/rhlI	**R1**
	lasR/rsaL/LasI	**L1**
P. aeruginosa PA7	rhlR/rhlI	**R1**
	lasR/rsaL/LasI	**L1**
P. aeruginosa PAO1	rhlR/rhlI	**R1**
	lasR/rsaL/LasI	**L1**
P. aeruginosa UCBPP-PA14	rhlR/rhlI	**R1**
	lasR/rsaL/LasI	**L1**
P. fuscovaginae UPB0736	sR/rsaM/sI	**M1**
	vR/rsaL/vI	**L1**
P. syringae 1448A	AhlR/AhlI	**R2**
P. syringae B728a	Psyr_1622/Psyr_1621	**R2**
P. syringae DC3000	psyR/psyI	**R2**
P. chlororaphis PCL1391	phzR/phzI	**R2**
P. flourescens 2-79	phzR/phzI	**R2**
P. flourescens NCIMB 10586	mupR/mupX/mupI	**X5**
P. putida WCS358	uR/rsaL/uI	**L1**
P. putida PCL1445	ppuR/rsaL/ppuI	**L1**
P. putida IsoF	ppuR/rsaL/ppuI	**L1**

6.5. Burkholdéria kromoszómák vizsgálata

6.5.1. Topológiák elhelyezkedése a kromoszómán

A *burkholdéria* rendbe tartozó baktériumok AHL alapú kommunikációval összefüggő génjeinek hasonlóság alapú csoportosulása nagyrészt megegyezik a taxonómiai alcsoportokkal. Következő lépésként a topológiák baktérium kromoszómán való elhelyezkedését vizsgáltam meg. Ehhez az 5.1.2 bekezdésben említett kromoszóma térképet rajzoltam a következő baktérium alcsoportokhoz:

Burkholderia cepacia complex (BCC): az elnevezés egy közeli rokonságban álló, humán patogén baktérium csoportra utal, melyet klinikai mintákból izoláltak. Jelen állás szerint 17 különböző faj tartozik bele, köztük a *B. cepacia*, *B. multivorans*, *B. cenocepacia*, *B. vietnamiensis*, *B. ambifaria*, *B. stabilis*, *B. dolosa*, *B. anthina* és *B. pyrrocina*. [70]

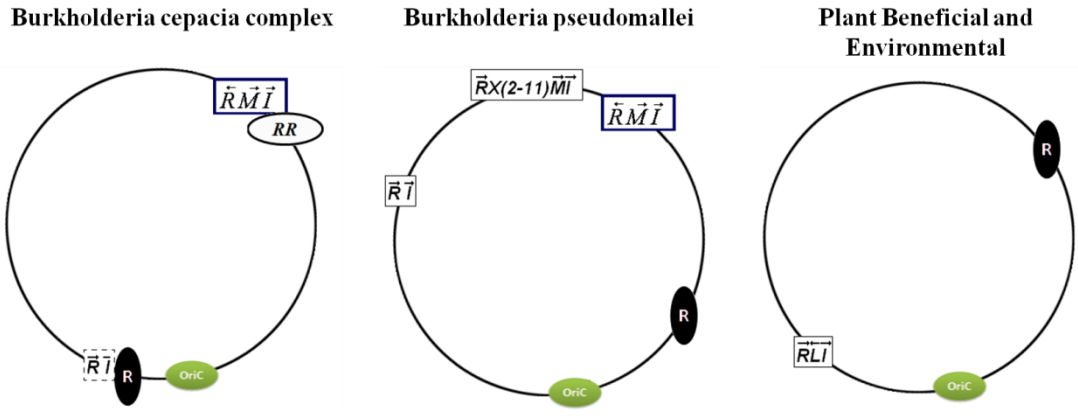
DOI:10.15774/PPKE.ITK.2014.006

Burkholderia pseudomallei csoport: két fő alcsoporttal rendelkezik: *B. mallei* és a *B. pseudomallei*. Mindkét alcsoportról ismert, hogy több AHL alapú *quorum sensing* található benne. [71]

Plant Beneficial and Environmental (PBE) csoport: ez a csoport 29 nem patogén fajt tartalmaz, amelyek a legtöbb esetben növényekkel állnak kapcsolatban. [72] Csupán néhány tagjánál találtak eddig *quorum sensing* rendszert, ami közeli rokonságban áll a *Pseudomonas aeruginosa* baktérium *LasR*/*LasI* rendszerével.

Mivel a bakteriális genomok cirkulárisak, ezért szükségem volt egy fix pontra, ami alapján az adott ábrákat azonos helyzetbe tudtam forgatni. Ennek a meghatározott pontnak a DNS replikáció kezdő helyét választottam (*origin of replication* = *OriC*), mivel ez minden bakteriális kromoszómán megtalálható. A térképeken ezt zöld oválissal jelöltem meg.

A 6.9. ábra mutatja meg a különböző csoportokhoz tartozó kromoszóma térképeket. A négyzetek jelölik a topológiákat, bennünk szerepeltetve az elrendeződés nyilas felírását. Fekete oválissal jelöltem a szóló vagy más néven árva *luxR* géneket. A **BCC** csoport ábráján két érdekes dolog is megfigyelhető. Először is az RR nevezető különleges topológia. Ez az olyan gén párokat jelöli, amelyek egymás mellett helyezkednek el a genomban, és mindkettő ugyanahhoz a fehérjecsaládhoz tartozik. Jelen esetben a *luxR* génekhez. Ezekre minden esetben igaz, hogy egy magas és egy alacsony valószínűségi pontszámmal rendelkező génpárról van szó. Az automata algoritmus ezeket is felismeri, de mivel csak kevés ilyen található a bakteriális genomokban, nem vizsgáltam őket részletesen. Mivel ebben a csoportban majdnem minden esetben megjelentek a genomban, az esetleges későbbi vizsgálatok miatt szerepeltettem őket az ábrán. A másik érdekesség a szaggatott keretű topológia. Ez a csoport egyes tagjainál megjelent, másoknál nem, ezért jelöltem eltérő módon.



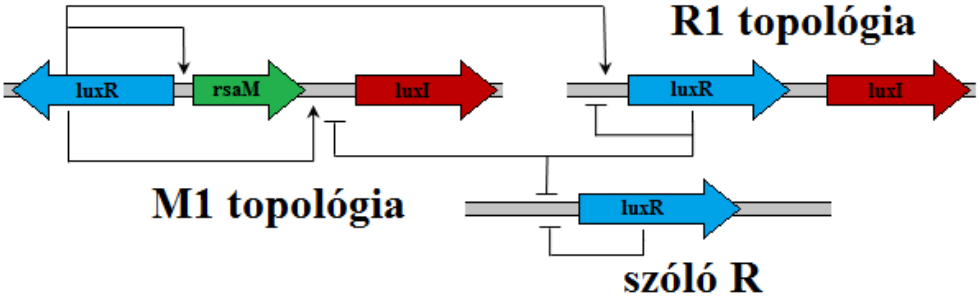
6.9. ábra Burkholdéria baktériumcsoportok kromoszóma térképe

DOI:10.15774/PPKE.ITK.2014.006

6.5.2. A topológiák egymásra gyakorolt hatása

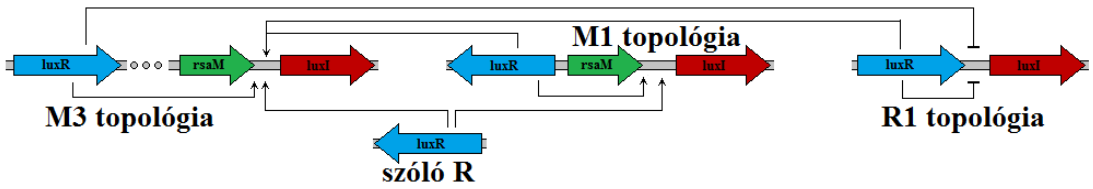
A *burkholdéria* kromoszómákon található különböző *quorum sensing* rendszerek nem függetlenek egymástól, hanem kölcsönösen kapcsolatban állnak egymással, így kialakítva egy komplex szabályozást. Ennek a szabályozásnak a felépítése baktérium csoportonként eltérő. Mindhárom csoportból választottam egy-egy baktériumot, amin bemutatom a *quorum sensing* rendszerek egymásra gyakorolt hatását. [73]

Burkholderia cenocepacia J2315 (BCC csoport): Az **M1** topológia *luxR* génje szabályozza a saját *luxI* génjét, és az **R1** topológia mindkettő génjét, amíg az **R1** topológia *luxR* génje negatívan szabályozza az **M1** topológia *luxI* génjét, ezzel létrehozva egy negatív szabályzó kört. A magányos *luxR* gén mind saját maga mind az **R1** topológia által negatívan van szabályozva. (6.10. ábra)



6.10. ábra A Burkholdéria cenocepacia J2315 komplex szabályzása

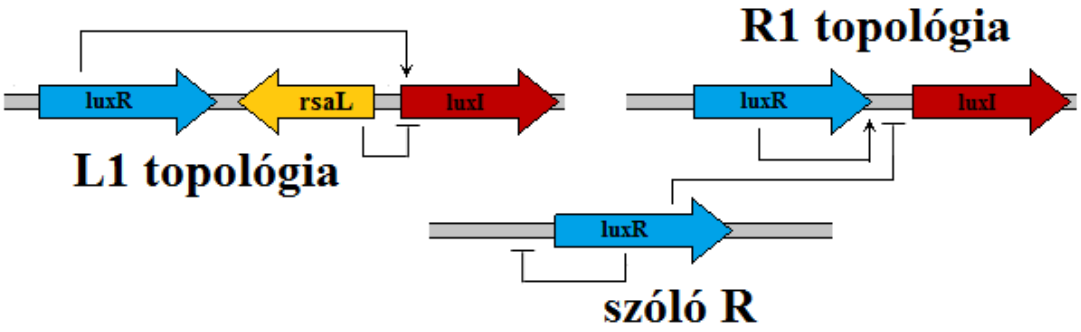
Burkholderia pseudomallei K92643 (pseudomallei group): Az **M3** topológia *luxR* génje szabályozza a saját *luxI* génjét, így létre hozva egy pozitív visszacsatolást. Ez a *luxI* gén ugyancsak szabályozva van az **R1** és **M3** topológiák *luxR* által. Az **R1** topológia *luxI* génje negatívan van szabályozva az **M3** topológia *luxR* génje által. A magányos *luxR* gén az **M1** topológia *luxI* génjének kifejtődését is szabályozza. Ugyanakkor elmondható, hogy minden *luxR* gén szabályozza a saját *luxI* génjét, de az **R1** topológia esetén negatívan. (6.11. ábra)



6.11. ábra A burkholderia pseudomallei K92643 komplex szabályozása

DOI:10.15774/PPKE.ITK.2014.006

Burkholderia xenovorans LB400 (PBE csoport): Ez a csoport viszonylag kevés *quorum sensing* topológiát tartalmaz, így a szabályozása sokkal egyszerűbb, mint az előző két esetben. A legtöbb csoportbeli baktérium csak egy topológiát tartalmaz (**L1**), de néhányan rendelkeznek egy **R1** topológiával is. A topológiák megszokott belső szabályozásain kívül csupán a magányos *luxR* gén fejt ki hatás az **R1** topológia *luxI* génjére. A két *quorum sensing* kör között nem tapasztalható kapcsolat. (6.12. ábra)

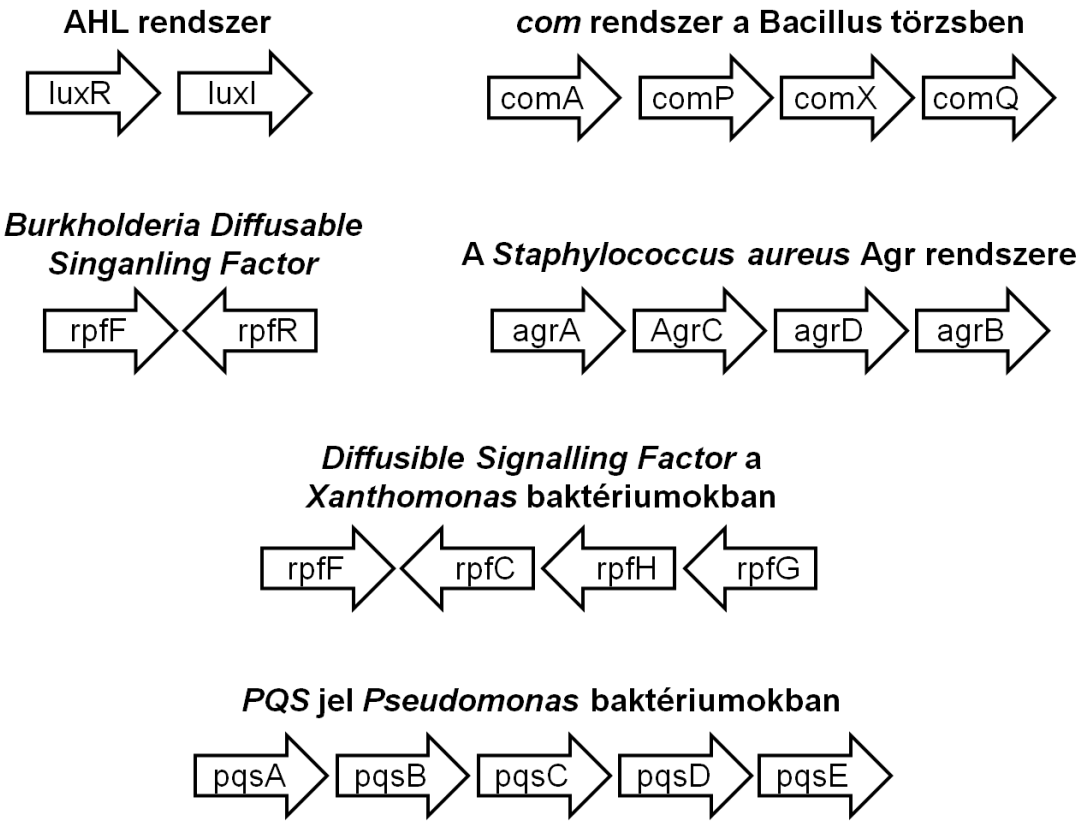


6.12. ábra A burkholderia xenovorans LB400 komplex szabályozása

DOI:10.15774/PPKE.ITK.2014.006

7. Az eredmények kiterjesztése más QS rendszerekre

A géntopológiák vizsgálatának rendszere elég általánosan van megfogalmazva, így az elkészült algoritmust alkalmazni lehet lényegében bármely kisméretű regulonra, abban az esetben, ha az azt alkotó gének fehérjetermékei elég nagyszámban ismertek ahhoz, hogy belőlük megfelelő minőségű HMM felismerőket lehessen építeni. Mivel a jelenleg ismert bakteriális genomokban nagyon sok az ismeretlen funkciójú gén, az ilyen jellegű vizsgálatoknak nagyon tág a tere. Munkatársaimmal elsősorban arra törekedtünk, hogy az általam kidolgozott elveket először a bakteriális kommunikáció - konkrétan a *quorum sensing* - egyéb rendszereire alkalmazzuk. Néhány ilyen rendszert az 7.1. ábra mutat be.



7.1. ábra Bakteriális kommunikációs rendszerek

Néhány bakteriális kommunikációs rendszer tipikus géntopológiája az irodalmi adatok alapján. Megjegyzendő, hogy az elsőnek listázott típus, az AHL rendszernek ebben a dolgozatban közölt vizsgálata mintegy 15 topológiai variánst talált a genomiális adatbázisokban, így feltehető, hogy a többi csoportban is lesznek újszerű elrendezések.

DOI:10.15774/PPKE.ITK.2014.006

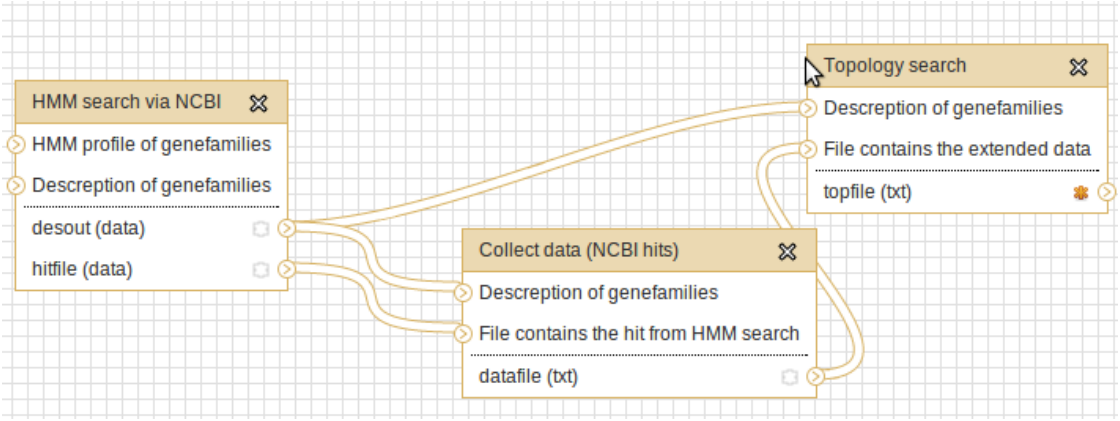
Az alkalmazás három fázisból állt, nevezetesen:

1) Az automatizált rendszer megtervezése és beprogramozása
2) A genomiális adatok analízise (keresés)
3) Az eredmények megjelenítése.

Ezeket részben Sonal Choudhary és Sanjarbek Hudaiberdiev PhD hallgatók végezték el, a dolgozatomban csak a főbb irányokat, illetve az általam végzett vagy tervezett fázisokat ismertetem.

1) A munkamenet beillesztése egy automatikus futtatást igénylő rendszerbe.

Erre a *Galaxy* [74] rendszert választottuk ki, amelyet általánosan használnak genom-elemzési feladatok automatizálására. A létrehozott rendszer vázlatát 7.2. ábra mutatja be.



7.2. ábra A kifejlesztett automatikus értékelő rendszer logikai vázlata.

Az így létrehozott rendszer bemeneti adatként HMM felismerőket fogad el, melyeket a felhasználónak kell előállítani. Ezt a megoldást azért választottam, mivel a HMM felismerők, illetve az azok alapjául szolgáló többszörös illesztések létrehozása az általános vélemény szerint egyébként is emberi beavatkozást igényel. Esetünkben ilyen feladat például a többdoménes fehérjék felismerőinek létrehozása, mint amilyen pl. az AHL rendszer *LuxR* fehérjéje.

2) A genomiális adatok analízise (keresés)

Előre kell bocsátani, hogy ez a fázis az adatbázisok nagyságától függ, továbbá azt is, hogy az előző pontban vázolt, kiterjesztett rendszer nemcsak komplett genomokat, hanem részben annotált, illetve nem annotált draft genomokat is vizsgál. Ez utóbbiak mennyisége pedig sokszorosan meghaladja a komplett genomokét, így az adatok végigkerestetésének igen nagy az időigénye. Néhány jellemző keresési időt a 7.1. táblázatban láthatunk.

DOI:10.15774/PPKE.ITK.2014.006

7.1. táblázat Az analízis keresési ideje

Keresési adatbázis	Fajok száma	Genomok összhossza	Szükséges idő*
Konkrét baktérium vizsgálata (*Pseudomonas aeruginosa PAO1*)	1	~6,3Mbp	~1 perc
Baktériumrend vizsgálata (Burkholdériák rendje)	~100	~600 Mbp	~30 perc
Összes teljes bakteriális genom (NCBI adatbázis)	~2650	~25ezer Mbp	~480 perc
Konkrét draft baktérium vizsgálata (*Eremococcus colicola ASC 139*)	1	~0,6 Mbp	~16 mp

* A keresési idő nagy mértékben függ az NCBI adatbázis leterheltségétől, mivel a futási idő nagy része az adatok lementésével telik el. A megadott értékek az átlagos időt jelzik.

3) Az eredmények prezentálása.

Az eredmények prezentálására és publikálására egy önálló honlap a legkézenfekvőbb, ezért Sonal Choudhary-val és Sanjarbek Hudaiberdiev-vel közösen olyan honlapot terveztünk, amely egyrészt tartalmazni fogja az eddig elemzett *quorum sensing* rendszerek összes ismert változatát, másrészt olyan keresésre is lehetőséget ad, amellyel egy felhasználó megvizsgálhatja, hogy az általa analizált bakteriális genomszekvencia tartalmazza-e az ismert *quorum sensing* rendszereket. Ez az adatbázis tehát nemcsak katalógus, hanem egy keresőmotort is tartalmaz, amelyet terveink szerint ki fogunk egészíteni a kapcsolódó legfőbb irodalmakkal is. Az elkészült rendszert a PPKE egyik szerver számítógépén tervezzük üzembe helyezni.

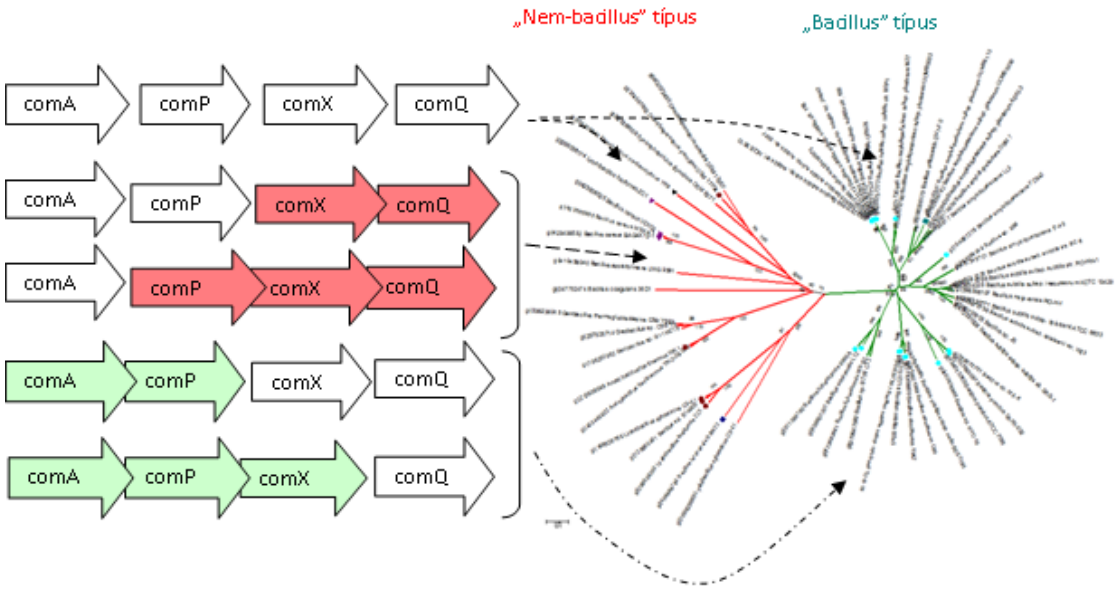
A web portál tervezett tartalma:

- A baktériumokban előforduló *quorum sensing* rendszerek részletes összefoglalása
- Eredmények szétbontásának lehetősége mind kommunikációs rendszertípus, mind genom típus (teljes, draft) alapján
- A kereső algoritmus elérése *Galaxy* rendszeren keresztül
- Statisztikák (átlapolási arány, GC tartalom)
- Részletes dokumentáció és tutoriálok

Végül felmerül a kérdés, hogy az AHL rendszernél talált általános konklúziók mennyire terjeszthetők ki más *quorum sensing* típusoknál. Az AHL rendszerrel kapcsolatos tapasztalatok röviden úgy foglalhatók össze, hogy az adott rendszereket alkotó gének többféle topológiai elrendezésben illetve regulációs mintázatok alapján tudják ellátni funkcióikat, és a fehérjék hasonlóságai a topológiai illetve regulációs csoportok szerint oszlanak meg. Ezt a kérdést meg kell vizsgálni majd az összes *quorum sensing* rendszeren. Az előzetes adatok

DOI:10.15774/PPKE.ITK.2014.006

jelenleg csak egy rendszernél; a *Bacillus Subtilis* peptid alapú jelzőrendszerénél állnak rendelkezésre. Ez a rendszer az AHL alapúhoz képes kevesebb topológiai variánst mutat, az esetek többségében a 4 gén ugyanis szorosan egymás után helyezkedik el. Előzetes vizsgálatok szerint azonban a konkrét topológián belül eltérő mintázatot mutat a gének átfedése (7.3. ábra), ami a génszabályozás különbözőségére utal. A gének hasonlósági megoszlásában, a filogenetikus fákban ugyanezeket a csoportokat látjuk, tehát itt is fenn áll az az általános konklúzió, hogy a gének a regulációs mintázat szerint oszlanak el a filogenetikus fán belül.



7.3. ábra Peptid alapú quorum sensing rendszer

A *Bacillus subtilis* baktériumban leirt peptid alapú *quorum sensing* jelzőrendszer géntopológiájának változata (balra), összehasonlítva az egyik gén (*comP*) filogenetikus fájával (jobbra). Látható, hogy a jellemző átfedésű topológiákban résztvevő gének a filogenetikus fán is külön csoportot alkotnak. (Előzetes eredmények, Sonal Choudhary és munkatársaitól)

DOI:10.15774/PPKE.ITK.2014.006

8. Konklúziók

Munkám a bakteriális kommunikáció egyik alaprendszere, az AHL alapú *quorum sensing* jelzőrendszer génjeinek feltérképezését tűzte ki célul. Ez a munka része a bakteriális kommunikáció általános feltérképezésének, amelyhez több, többé-kevésbé ismert jelzőrendszer is tartozik. Az AHL rendszer csak a Gram-negatív baktériumoknál, ezen belül a proteobaktériumok csoportjánál ismert. A munkám egyik célja ennek az általánosan ismert ténynek az ellenőrzése volt. Habár elég tág keresési beállításokat használtam, a proteobaktériumokon kívül nem sikerült teljes AHL rendszert találnom. Ezzel együtt természetesen nem zárható ki, hogy ilyen rendszer felbukkanhat a jövőben, hiszen a genomiális információ szintjén jelenleg csak elenyészően kis részét ismerjük a természetben előforduló baktériumoknak. Ezen kívül az általam használt teljes bakteriális genomok adatbázisa szükségszerűen „elfogult", hiszen a gyakorlati szempontból érdekes organizmusokra koncentrál, ezért nem is tekinthető reprezentatívnak a természetben megtalálható összes baktériumra nézve.

A vizsgált AHL *quorum sensing* gének topológiáiról észrevettem, hogy néhány egyszerű szabályt követnek, bár topológiájuk még mindig változatosabb, mint azt korábban sejteni lehetett. A gének klaszterezése azt mutatta, hogy az elrendeződés szoros összefüggésben van a szabályozással, és a rendszer génjei együtt fejlődtek. Megfigyeltem, hogy különböző fajok azonos elrendeződésű *quorum sensing* génjei jobban hasonlítanak egymásra, mint az ugyanabban a fajban előforduló azonos funkciójú génekre. A topológiák jól megfigyelhető, konzervált mintákat mutatnak, amelyek csoportosulása nem véletlenszerű a taxonómiai osztályokban. Ez persze nem azt jelenti, hogy szigorú összefüggéseket tudnánk vonni a konzervált topológiák és a funkciók között, de a topológiai minták egy véges halmazának kifejeződése egy stabil, pozitív autóregulációs kört eredményez.

A talált topológiai elrendezéseket két csoportra osztottam. Egyszerű topológiáknak azokat neveztem, ahol csak az autoindukciós kör két alapgénje, a jelet előállító szintáz enzim és a jelet érzékelő regulátor gén szerepel. Összetett topológiáknak azokat hívtam, ahol e két gén között és mellett több más gén is jelen van. Az egyszerű topológiai elrendezéseknek minden elképzelhető orientációs változata előfordul a vizsgált adatbázisokban. Külön érdekesek a két alapgén között elhelyezkedő, minden bizonnyal szorosan együtt szabályozott gének. Ezek az ismert esetek nagy többségében negatív regulációs szerepet töltenek be, tehát stabilizálják az autoindukciós visszacsatolás működését: megakadályozzák, hogy a rendszer aktivitása minden határon túl nőhessen. Az esetek kis többségében úgy nevezett DNS mobilizációs géneket találtunk a két alapgén között. Ezek pontos szerepe nem ismert, az viszont igen, hogy egyes – közelebbről itt nem vizsgált – baktériumoknál a *quorum sensing*

DOI:10.15774/PPKE.ITK.2014.006

rendszer valóban DNS mobilizációt, konkrétan plazmid átadást szabályoz. Ezáltal a plazmid átadása – egy energetikailag költséges folyamat – akkor megy csak végbe, ha a donorsejt jeltermelő sejtekkel van körülvéve. Valószínűnek tartom tehát, hogy ezek a DNS mobilizáló gének az AHL rendszeren belül ezzel némileg analóg szerepet tölthetnek be. A bonyolult topológiai elrendezéseket nem vizsgáltam részletesen, ezek között több jól ismert és sokat vizsgált rendszert fedezhetünk fel, mint a rhizóbiumok és az *Agrobaktérium tumafaciens*. Megvizsgáltam a topológiai elrendezéseket a *Pseudomonas* és a *Burkholderia* törzsek ismert teljes genomjain belül. Azt találtam, hogy a legtöbb topológia típus előfordul mind a két csoportban, tehát feltehető, hogy a topológiai elrendezés önmaga viszonylag könnyen változik, feltéve, hogy a szabályozási tulajdonságok (pl. a jelre való reagálás, a válasz stabilitása) megmarad. Látszólag ezek a regulációs kritériumok többféleképpen is kielégíthetők, de a kialakult topológiák jól követik a törzsön belüli eloszlást, például a *Burkholderia* törzsön belül az ismert alcsoportok (*cenocepacia, mallei* illetve növényi szimbióta) az AHL típusú *quorum sensing* gének topológiája alapján is elkülönülnek egymástól. . Az ilyen jellegű különbségeket a biológiában az ortológ-paralóg fogalompár írja le [75, 76]. Eredményeim arra mutatnak, hogy az eddig egységes ortológ csoportnak gondolt *LuxR* (vagy *LuxI*) fehérjecsalád tovább osztható a gének lokális elrendeződése alapján. Ez a jelenség általánosnak tűnik, a felosztás egyformán fennáll a *luxR* és *luxI* génekre, sőt – jelenleg folyó vizsgálatok szerint – a *LuxR* fehérje doménjeire is. Vagyis az egyes, feltehetően különböző feladatokat ellátó kommunikációs fehérjék önálló orológ csoportéként fejlődnek a természetben. Végül egy esetben, a *Burkholderi*a törzs tagjainál vizsgáltam a kromoszómán belüli eloszlást, itt is látszottak típusok, de nem lehetett egyértelmű szabályszerűségeket megfigyelni.

Az automatizált kereső algoritmus nagyban megkönnyítette a bakteriális genomok *quorum sensing*gel kapcsolatos génjeinek vizsgálatát, mert a rendelkezésre álló óriási mennyiségű adatból kiszűrte a valóban fontosakat, így már csak egy jóval kisebb adathalmazt kellett vizsgálni. Ezen kívül lehetővé tette az információk naprakészen tartását, így folyamatosan növelve az elemzés alatt álló gének számát.

Fontos megjegyezni, hogy az adatbázisok folyamatos, antomatikus adatbányászatának a jövőben várhatóan egyre nagyobb jelentősége lesz. A baktériumok génállománya például az általános felfogás szerint igen könnyen feltérképezhető, de az ezzel foglalkozó szakemberek tudják, hogy ez csak a géneknek tipikusan csak mintegy felét kitevő „törzsállományra” igaz, az ezen felüli gének funkciója még ebben az egyszerűnek tartott esetben is többnyire ismeretlen marad. Vagyis még ebben a jólismert csoportban is sok a felfedezetlen génfunkció, melyet az adatbázisok „hipotetikus” jelzővel szoktak jelezni. Az általam fejlesztett kereső-rendszer az AHL géneket analizálta, és ebben a szűk csoportban is több mint 50 hipotetikusnak jelzett gén funkcióját sikerült megállapítani.

DOI:10.15774/PPKE.ITK.2014.006

Dolgozatom lezárásakor ez a munka még folytatódik, jelenleg terjesztjük ki a vizsgálatokat más bakteriális jelzőrendszerekre. Olyan automatikus adatbányászati rendszert fejlesztünk ki, melyekkel a bakteriális *quorum sensing* génjei folyamatosan naprakészen tarthatók. Közismert, hogy a genomiális adatok egyre növekvő hányada annotáció nélkül marad, így szinte biztosra vehető, hogy az automatikus értelmező rendszereknek a jövőben jelentős szerepe lesz.

DOI:10.15774/PPKE.ITK.2014.006

9. Publikációk

Zsolt Gelencsér, Borisz Galbáts, Juan F. Gonzalez, K. Sonal Choudhary, Sanjarbek Hudaiberdiev, Vittorio Venturi, and Sándor Pongor “Chromosomal Arrangement of AHL-Driven Quorum Sensing Circuits in Pseudomonas” ISRN Microbiology, vol. 2012, Article ID 484176, 6 pages, 2012.

Dóra Bihary, Ádám Kerényi, **Zsolt Gelencsér**, Sergiu Netotea, Attila Kertész-Farkas, Vittorio Venturi, Sándor Pongor "Simulation of communication and cooperation in multispecies bacterial communities with an agent based model" Scalable Computing: Practice and Experience Volume 13, Number 1, pp. 21–28.

Zsolt Gelencsér, Kumari Sonal Choudhary, Bruna Goncalves Coutinho, Sanjarbek Hudaiberdiev, Borisz Galbáts, Vittorio Venturi, and Sándor Pongor "Classifying the Topology of AHL-Driven Quorum Sensing Circuits in Proteobacterial Genomes" Sensors, vol. 12(5), pp. 5432-5444, 2012.

Kumari Sonal Choudhary, Sanjarbek Hudaiberdiev, **Zsolt Gelencsér**, Bruna GonçalvesCoutinho, Vittorio Venturi, and Sándor Pongor "The Organization of the Quorum Sensing luxI/R Family Genes in Burkholderia," Int J Mol Sci, vol. 14, pp. 13727-13747, 2013.

10. Referenciák

[1] J. Maynard Smith and E. Szathmary, *The Major Transitions in Evolution*. Oxford, England: Oxford University Press, 1995.

[2] K. H. Nealson, *et al.*, "Cellular control of the synthesis and activity of the bacterial luminescent system," *J Bacteriol*, vol. 104, pp. 313-22, Oct 1970.

[3] A. Eberhard, *et al.*, "Structural identification of autoinducer of Photobacterium fischeri luciferase," *Biochemistry*, vol. 20, pp. 2444-9, Apr 28 1981.

[4] J. Engebrecht, *et al.*, "Bacterial bioluminescence: isolation and genetic analysis of functions from Vibrio fischeri," *Cell*, vol. 32, pp. 773-81, Mar 1983.

[5] W. C. Fuqua and S. C. Winans, "A LuxR-LuxI type regulatory system activates Agrobacterium Ti plasmid conjugal transfer in the presence of a plant tumor metabolite," *J Bacteriol*, vol. 176, pp. 2796-806, May 1994.

[6] C. Fuqua, *et al.*, "Regulation of gene expression by cell-to-cell communication: acyl-homoserine lactone quorum sensing," *Annu Rev Genet*, vol. 35, pp. 439-68, 2001.

[7] W. C. Fuqua, *et al.*, "Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators," *J Bacteriol*, vol. 176, pp. 269-75, Jan 1994.

[8] C. M. Waters and B. L. Bassler, "Quorum sensing: cell-to-cell communication in bacteria," *Annu Rev Cell Dev Biol*, vol. 21, pp. 319-46, 2005.

[9] N. A. Whitehead, *et al.*, "Quorum-sensing in Gram-negative bacteria," *FEMS Microbiol Rev*, vol. 25, pp. 365-404, Aug 2001.

[10] M. I. More, *et al.*, "Enzymatic synthesis of a quorum-sensing autoinducer through use of defined substrates," *Science*, vol. 272, pp. 1655-8, Jun 14 1996.

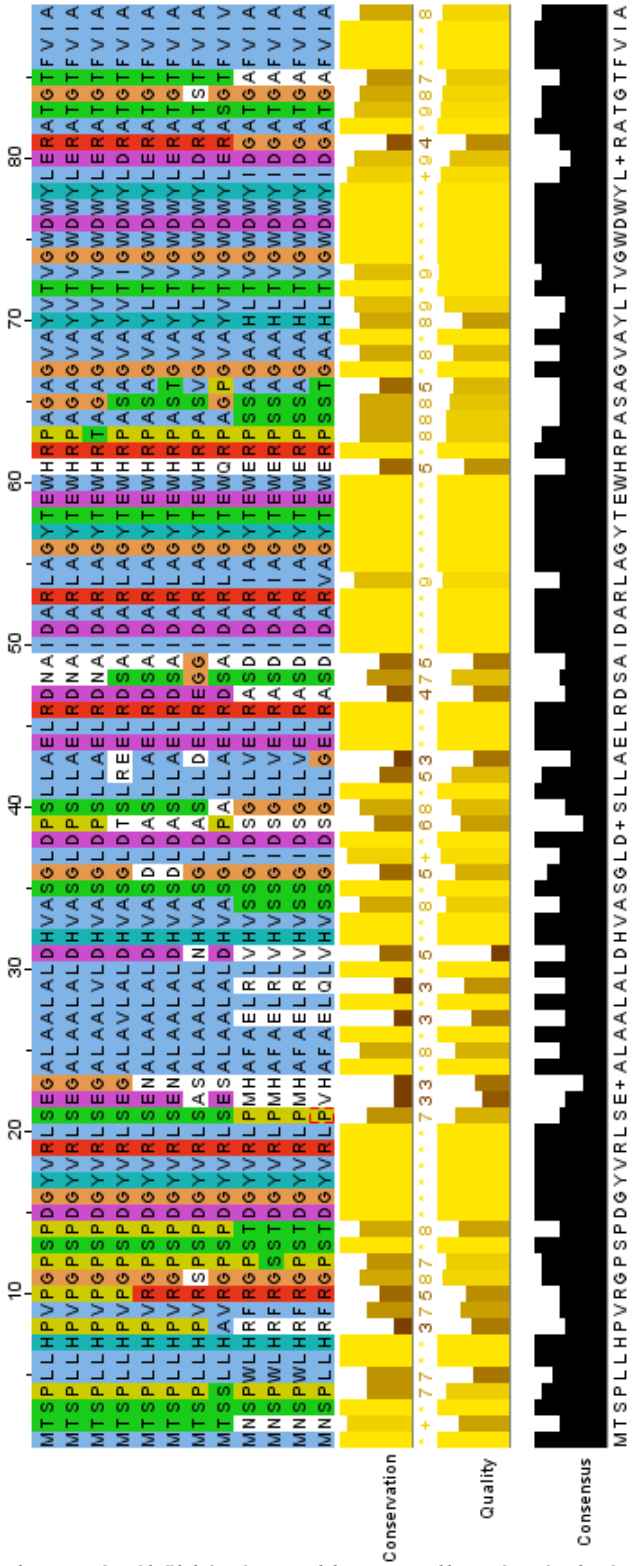
[11] B. L. Hanzelka and E. P. Greenberg, "Evidence that the N-terminal region of the Vibrio fischeri LuxR protein constitutes an autoinducer-binding domain," *J Bacteriol*, vol. 177, pp. 815-7, Feb 1995.
[12] M. Welch, *et al.*, "N-acyl homoserine lactone binding to the CarR receptor determines quorum-sensing specificity in Erwinia," *EMBO J*, vol. 19, pp. 631-41, Feb 15 2000.
[13] V. Venturi, *et al.*, "The virtue of temperance: built-in negative regulators of quorum sensing in Pseudomonas," *Mol Microbiol*, vol. 82, pp. 1060-70, Dec 2011.
[14] M. Mattiuzzo, *et al.*, "The plant pathogen Pseudomonas fuscovaginae contains two conserved quorum sensing systems involved in virulence and negatively regulated by RsaL and the novel regulator RsaM," *Environ Microbiol*, vol. 13, pp. 145-62, Jan 2011.
[15] A. M. Stevens, *et al.*, "Mechanisms and synthetic modulators of AHL-dependent gene regulation," *Chem Rev*, vol. 111, pp. 4-27, Jan 12 2011.
[16] R. D. Fleischmann, *et al.*, "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science*, vol. 269, pp. 496-512, Jul 28 1995.
[17] "Genome online database (gold) website: http://genomesonline.org."
[18] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase," *J Mol Biol*, vol. 94, pp. 441-8, May 25 1975.
[19] R. Cullum, *et al.*, "The next generation: using new sequencing technologies to analyse gene regulation," *Respirology*, vol. 16, pp. 210-22, Feb 2011.
[20] E. R. Mardis, "A decade's perspective on DNA sequencing technology," *Nature*, vol. 470, pp. 198-203, Feb 10 2011.
[21] "http://www.genome.gov/sequencingcosts."
[22] "NCBI GenBank Statictics 2008: http://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008/."
[23] R. Overbeek, *et al.*, "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes," *Nucleic Acids Res*, vol. 33, pp. 5691-702, 2005.
[24] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J Mol Biol*, vol. 48, pp. 443-53, Mar 1970.
[25] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J Mol Biol*, vol. 147, pp. 195-7, Mar 25 1981.
[26] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proc Natl Acad Sci U S A*, vol. 89, pp. 10915-9, Nov 15 1992.
[27] J. D. Thompson, *et al.*, "Multiple sequence alignment using ClustalW and ClustalX," *Curr Protoc Bioinformatics*, vol. Chapter 2, p. Unit 2 3, Aug 2002.
[28] M. Clamp, *et al.*, "The Jalview Java alignment editor," *Bioinformatics*, vol. 20, pp. 426-7, Feb 12 2004.
[29] S. R. Eddy, "What is a hidden Markov model?," *Nat Biotechnol*, vol. 22, pp. 1315-6, Oct 2004.
[30] S. R. Eddy, "Hidden Markov models," *Curr Opin Struct Biol*, vol. 6, pp. 361-5, Jun 1996.
[31] S. R. Eddy, "A new generation of homology search tools based on probabilistic inference," *Genome Inform*, vol. 23, pp. 205-11, Oct 2009.
[32] S. F. Altschul, *et al.*, "Basic local alignment search tool," *J Mol Biol*, vol. 215, pp. 403-10, Oct 5 1990.
[33] R. L. Tatusov, *et al.*, "A genomic perspective on protein families," *Science*, vol. 278, pp. 631-7, Oct 24 1997.
[34] A. Bairoch, "PROSITE: a dictionary of sites and patterns in proteins," *Nucleic Acids Res*, vol. 19 Suppl, pp. 2241-5, Apr 25 1991.
[35] S. Dhir, *et al.*, "Detecting atypical examples of known domain types by sequence similarity searching: the SBASE domain library approach," *Curr Protein Pept Sci*, vol. 11, pp. 538-49, Nov 2010.
[36] D. A. Benson, *et al.*, "GenBank," *Nucleic Acids Res*, vol. 28, pp. 15-8, Jan 1 2000.

[37] G. Cochrane, *et al.*, "EMBL Nucleotide Sequence Database: developments in 2005," *Nucleic Acids Res*, vol. 34, pp. D10-5, Jan 1 2006.
[38] K. Okubo, *et al.*, "DDBJ in preparation for overview of research activities behind data submissions," *Nucleic Acids Res*, vol. 34, pp. D6-9, Jan 1 2006.
[39] C. H. Wu, *et al.*, "The Universal Protein Resource (UniProt): an expanding universe of protein information," *Nucleic Acids Res*, vol. 34, pp. D187-91, Jan 1 2006.
[40] C. O'Donovan, *et al.*, "High-quality protein knowledge resource: SWISS-PROT and TrEMBL," *Brief Bioinform*, vol. 3, pp. 275-84, Sep 2002.
[41] T. N. Bhat, *et al.*, "The PDB data uniformity project," *Nucleic Acids Res*, vol. 29, pp. 214-8, Jan 1 2001.
[42] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res*, vol. 28, pp. 27-30, Jan 1 2000.
[43] R. L. Tatusov, *et al.*, "The COG database: an updated version includes eukaryotes," *BMC Bioinformatics*, vol. 4, p. 41, Sep 11 2003.
[44] M. Ashburner, *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-9, May 2000.
[45] "COG honlapja: http://www.ncbi.nlm.nih.gov/COG/."
[46] "NCBI hivatalos honlap: http://www.ncbi.nlm.nih.gov/."
[47] "NCBI FTP szervere: ftp.ncbi.nlm.nih.gov."
[48] "UniProt adatbázis: http://www.uniprot.org/."
[49] J. E. Stajich, *et al.*, "The Bioperl toolkit: Perl modules for the life sciences," *Genome Res*, vol. 12, pp. 1611-8, Oct 2002.
[50] "BioPerl hivatalos honlapja: http://www.bioperl.org."
[51] "BioPython hivatalos honlapja: http://www.biopython.org."
[52] "PHYLIP hivatalos oldala: http://evolution.genetics.washington.edu/phylip.html."
[53] K. Rutherford, *et al.*, "Artemis: sequence visualization and annotation," *Bioinformatics*, vol. 16, pp. 944-5, Oct 2000.
[54] "Artemis: http://www.sanger.ac.uk/resources/software/artemis/."
[55] "HMMER hivatalos honlapja: http://hmmer.janelia.org/."
[56] S. Subramoni and V. Venturi, "LuxR-family 'solos': bachelor sensors/regulators of signalling molecules," *Microbiology*, vol. 155, pp. 1377-85, May 2009.
[57] Y. Lequette, *et al.*, "A distinct QscR regulon in the Pseudomonas aeruginosa quorum-sensing circuit," *J Bacteriol*, vol. 188, pp. 3365-70, May 2006.
[58] R. J. Case, *et al.*, "AHL-driven quorum-sensing circuits: their frequency and function among the Proteobacteria," *ISME J*, vol. 2, pp. 345-9, Apr 2008.
[59] A. B. Goryachev, "Understanding bacterial cell-cell communication with computational modeling," *Chem Rev*, vol. 111, pp. 238-50, Jan 12 2011.
[60] A. B. Goryachev, "Design principles of the bacterial quorum sensing gene networks," *Wiley Interdiscip Rev Syst Biol Med*, vol. 1, pp. 45-60, Jul-Aug 2009.
[61] D. I. Sharif, *et al.*, "Quorum sensing in Cyanobacteria: N-octanoyl-homoserine lactone release and response, by the epilithic colonial cyanobacterium Gloeothece PCC6909," *ISME J*, vol. 2, pp. 1171-82, Dec 2008.
[62] L. Falquet, *et al.*, "The PROSITE database, its status in 2002," *Nucleic Acids Res*, vol. 30, pp. 235-8, Jan 1 2002.
[63] M. Sanchez-Contreras, *et al.*, "Quorum-sensing regulation in rhizobia and its role in symbiotic interactions with legumes," *Philos Trans R Soc Lond B Biol Sci*, vol. 362, pp. 1149-63, Jul 29 2007.
[64] C. E. White and S. C. Winans, "Cell-cell communication in the plant pathogen Agrobacterium tumefaciens," *Philos Trans R Soc Lond B Biol Sci*, vol. 362, pp. 1135-48, Jul 29 2007.
[65] G. Rampioni, *et al.*, "The Pseudomonas quorum-sensing regulator RsaL belongs to the tetrahelical superclass of H-T-H proteins," *J Bacteriol*, vol. 189, pp. 1922-30, Mar 2007.
[66] Z. Gelencsér, *et al.*, "Chromosomal arrangement of AHL-driven quorumsensing circuits in Pseudomonas.," *ISRN Microbiol.*, p. 484176, 2012.

[67] C. S. Tsai and S. C. Winans, "LuxR-type quorum-sensing regulators that are detached from common scents," *Mol Microbiol*, vol. 77, pp. 1072-82, Sep 2010.
[68] D. C. Krakauer, "Stability and evolution of overlapping genes," *Evolution*, vol. 54, pp. 731-9, Jun 2000.
[69] H. Hirakawa, *et al.*, "Activity of the Rhodopseudomonas palustris p-coumaroyl-homoserine lactone-responsive transcription factor RpaR," *J Bacteriol*, vol. 193, pp. 2598-607, May 2011.
[70] V. Venturi, *et al.*, "Quorum sensing in the Burkholderia cepacia complex," *Res Microbiol*, vol. 155, pp. 238-44, May 2004.
[71] J. C. Larsen and N. H. Johnson, "Pathogenesis of Burkholderia pseudomallei and Burkholderia mallei," *Mil Med*, vol. 174, pp. 647-51, Jun 2009.
[72] Z. R. Suarez-Moreno, *et al.*, "Commonalities and differences in regulation of N-acyl homoserine lactone quorum sensing in the beneficial plant-associated burkholderia species cluster," *Appl Environ Microbiol*, vol. 76, pp. 4302-17, Jul 2010.
[73] K. S. Choudhary, *et al.*, "The Organization of the Quorum Sensing luxI/R Family Genes in Burkholderia," *Int J Mol Sci*, vol. 14, pp. 13727-47, 2013.
[74] "Galaxy hivatalos honlap: http://galaxyproject.org/."
[75] A. Kuzniar, *et al.*, "The quest for orthologs: finding the corresponding gene across genomes," *Trends Genet*, vol. 24, pp. 539-51, Nov 2008.
[76] T. Gabaldon and E. V. Koonin, "Functional and evolutionary implications of gene orthology," *Nat Rev Genet*, vol. 14, pp. 360-6, May 2013.

DOI:10.15774/PPKE.ITK.2014.006

I. Melléklet



Az RsaM gének egy részéből készített többszörös illesztés részletének megjelenítése Jalview program segítségével. A hisztogramok az adott pozíciók konzerváltságát, minőségét és konszenzusát mutatják.

DOI:10.15774/PPKE.ITK.2014.006

II. Melléklet

```
>gi|15596624|ref|NP_250118.1| hypothetical protein PA1427 [Pseudomonas aeruginosa PAO1]
MVVEQVVVSETAPVDDPAGRVEHEVFQVFALEGDLAGEKTALGGLRQQMAAVAQEQRLSRCQRHLEQSLG
GIPELQIDQQAVVRGVMEDLQALAPFGQFEAAQAVVGQIPELEPGVGAGDGQALAVEKAQVPHRWTMQEP
PIFAQGEVFQDRQRCFALCRGRVEGPVEGQAGAGQVVQHQWGVDRKG
>gi|15596625|ref|NP_250119.1| hypothetical protein PA1428 [Pseudomonas aeruginosa PAO1]
MHIRKRVDADNTALVDLWERSVRATHDFLSEADIVELYPQVRDLYLPAVEVWVLVDDDGVAQGFIGLNQA
HVEMLFVEPGLRGRGIGRRLLDHARATWPRLSVDVNEQNPQACGFYRHYGFRQTGRSATDSAGRPFPLLH
MSL
>gi|15596626|ref|NP_250120.1| cation-transporting P-type ATPase [Pseudomonas aeruginosa PAO1]
MNGIPPLSSRDAHASPPQRVLERLHSSAAGLDADEAARRLAAHGYNRLPAPKRQGPLLRLLRQFHNVLLY
MMLFASLVTALLGFWVDSAVILLAVVVNALIGFVQEGKAANALDAIRDMLSLHALVLRDGQRQALDAERL
VPGDVVLLASGDRVPADLRLFETKNFHVDESALTGESVPVEKGCVAVAIDALLGDRRCMAYSGTLVTSGQ
ARGVVVATAGDTELGRIGTLLREVRTLATPLLRQIASFSRWLALAILLLAGATFVLGTLWQGQPMVDMFM
LVVALTASAIPEGLPAIMTVILALGVQRMARRNAIVSRLPAVETLGSVTVICSDKTGTLTRNEMTVQRIV
TADQVIEVSGAGYAPLGGFSHNGEGLDPAGRDDLQEIGRAALLCNEARLHQEGEAWQLEGDPTEGALLSL
GLKLGLDPQALAAERPRSDAIPFESEHRFMATLHHDHAGQAMVYLKGAPERILDMCEAERVGDSVRPLDP
DYWRRLATDTAARGLRLLAIARRAMPAEQRTLDFADVEHGFTLLALVGIIDPPREEAVAAVAECQAAGIA
VKMITGDHVDTARAIGAMLGIGIDRPALTGAEIELLDDQRLREVLPGVDVFARASPEHKLRLVQALQASG
EVVAMTGDGVNDAPALKRADVGVAMGNKGTEAAKEAAEVVLADDNFATIANAVREGRAVYDNLKKFILFM
LPTNGGEALIVIAAILFQLTLPMTPAQILWINMVTSSTLGLALAFDPAERGLMQRPPRPPAEPLLSLFFV
WRVLLVSLLMMAGALGLFLWELEHGTGLESARTMAVNSVVVCEMFYLLNSRHIYDSVLSREGLFGNRQVL
LAIAACVMLQLLYTYAPPLQALFGSVGLAPGEWARVLLAGLGLFCVAELEKWLCRRVRARQA
>gi|15596627|ref|NP_250121.1| transcriptional regulator LasR [Pseudomonas aeruginosa PAO1]
MALVDGFLELERSSGKLEWSAILQKMASDLGFSKILFGLLPKDSQDYENAFIVGNYPAAWREHYDRAGYA
RVDPTVSHCTQSVLPIFWEPSIYQTRKQHEFFEEASAAGLVYGLTMPLHGARGELGALSLSVEAENRAEA
NRFMESVLPTLWMLKDYALQSGAGLAFEHPVSKPVVLTSREKEVLQWCAIGKTSWEISVICNCSEANVNF
HMGNIRRKFGVTSRRVAAIMAVNLGLITL
>gi|15596628|ref|NP_250122.1| regulatory protein RsaL [Pseudomonas aeruginosa PAO1]
MASHERTQPQNMAFRAKATRTARRESQETFWSRFGISQSCGSRFENGENLPFPIYLLLHFYIEGQITDRQ
LADLRGKIRE
>gi|15596629|ref|NP_250123.1| autoinducer synthesis protein LasI [Pseudomonas aeruginosa PAO1]
MIVQIGRREEFDKKLLGEMHKLRAQVFKERKGWDVSVIDEMEIDGYDALSPYYMLIQEDTPEAQVFGCWR
ILDTTGPYMLKNTFPELLHGKEAPCSPHIWELSRFAINSGQKGSLGFSDCTLEAMRALARYSLQNDIQTL
VTVTTVGVEKMMIRAGLDVSRFGPHLKIGIERAVALRIELNAKTQIALYGGVLVEQRLAVS
>gi|15596630|ref|NP_250124.1| hypothetical protein PA1433 [Pseudomonas aeruginosa PAO1]
MSLLKQLFLAICLFLVVAFSGSFVSSVENSREQLRGQLRSHAQDAATALGLSLTPHVDDPAMVQLMVSSI
FDSGYFASIRVIDIKSGKPLVERVQAHAERTVPGWFERLVDLQPQGGDALIMRGWEQAARVEVVSHPQFA
LARLWDSALGSLYWLLACGAASLLLGGWLLRRQLRPLDQMVRQAHAISRREFLSLPRLPRTPELRRVVQA
MNQMVEKLRTLFAEEAARSDKLRAQAYQDSLTGLPNRRLFDARLNEQLGAGEHEHAGQLLLLRLNDLNGL
NQRLGGQRTDELIQAVARLLVDSCGQQGRADWLLARSRGGEFAVLAPGCSREQAERLAEELCEGLENLAR
TGASDLTPVAYLGISAFAEGDSPADLLARADQALAQAESQPAQPWASQDGTALAALNDSQDWHDWIDQAL
TERRLLLYFQPVVDCLDTQRVLHHKVLARLLDPQATAIAAGRFLPWIERFGWAARMDLAMLEQSLEHLRR
HPRPLALSLSAASVRNAQTFAPLLALLKAHPQEARQLTLELDERHLPAAAELERLSQVLRELGCGLGLQH
FGGRFSLIGNLTHLGLAYLKLDGCYLHAVDREGDKRLFIEAVYRTTHSIDLPLIAEQVETLGELEVLREM
GLRGAMGRLFGSPAPWSGDA
>gi|15596631|ref|NP_250125.1| hypothetical protein PA1434 [Pseudomonas aeruginosa PAO1]
MSPTPGARRPCPRAYAPWLLSLAATLLLAVGAALAQWDLESILSRAEQRYGELGAAKSRLGDWGRLLEQG
GTLDEAAKLRAVNDFFNRSLRFTDDIEIWQQEDYWATPVEALVKGAADCEDYAIAKYVTLRRLGVASDKL
RITYVKALRLNQAHMVLTWYASPGADPLVLDNLIGEIRPASQRDDLLPVYAFNAEGLWLPGADGGRRTGD
SKKLSRWQDLLTKMRAEGLDLDAPKED
>gi|15596632|ref|NP_250126.1| Resistance-Nodulation-cell Division (RND) [Pseudomonas aeruginosa PAO1]
MQALRSGGGRVLVGVLAAGLVAFGGWAWLGGDAGAKAAPAPARVPVIVARVERRDVEQQVSGIGTVTSLH
NVVIRTQIDGQLTRLLVSEGQMVEAGELLATIDDRAVVAALEQAQASRASNQAQLKSAEQDLQRYRSLYA
ERAVSRQLLDQQQATVDQLRATLKANDATINAERVRLSYTRITSPVSGKVGIRNVDVGNLVRVGDSLGLF
SVTQIAPISVVFSLQQEQLLQLQALLGGEAAVRAYSRDGGSALGEGRLLTIDNQIDSSTGTIRVRASFDN
RQARLWPGQFVAVSLHTGVRRDQLVLSSKAVRRGLEGNFVYRVADDRVEAVPVRVLQDIDGLSVVEGLAS
GDQVVVDGHSRLMPGALVDIQEPRPSLAQATERRP
```

A Pseudomonas aeruginosa nevű baktérium NC_002516 nevű faa fájl részlete. Jól láthatóak a FASTA formátumra jellemző, '>' jellel bevezetett fejléc sorok, melyek legelső szava az adott szekvencia azonosítója. Jelen esetben egy összetett azonosítót láthatunk, amely a GI és RefSeq számokat tartalmazza. A pirossal kiemelt rész egy LasR-RsaL-LasI fehérje hármas.

DOI:10.15774/PPKE.ITK.2014.006

III. Melléklet

```
CGCCGGCGCAGATCCTCTGGATCAACATGGTCACCTCCAGCACCCTGGGCCTGGCGCTGGCCTTCGATCC
GGCCGAGCGCGGCCTGATGCAGCGTCCGCCGCGTCCGCCGGCGGAGCCCCTGTTGTCGCTGTTCTTCGTC
TGGCGGGTGCTGCTGGTCTCGCTGCTGATGATGGCGGGCGCCCTGGGCCTGTTCCTCTGGGAGCTGGAGC
ATGGCACCGGGCTGGAGAGCGCGCGGACCATGGCGGTGAACAGCGTGGTGGTGTGCGAGATGTTCTACCT
GCTCAACAGCCGGCATATCTACGATTCGGTGCTCAGCCGCGAGGGCCTGTTCGGCAACCGCCAGGTCTTG
CTGGCGATCGCCGCCTGCGTGATGCTGCAACTGCTCTATACCTATGCGCCGCCGTTGCAGGCGCTGTTCG
GCTCGGTCGGCCTGGCCCCGGGCGAGTGGGCGCGGGTGCTGCTGGCCGGGCTCGGCCTGTTCTGCGTCGC
CGAACTGGAAAAGTGGCTATGTCGCCGGGTCCGTGCCCGGCAGGCCTGA
>gi|110645304|ref|NC_002516.2|:1558171-1558890 Pseudomonas aeruginosa PAO1 chromosome, complete genome
ATGGCCTTGGTTGACGGTTTTCTTGAGCTGGAACGCTCAAGTGGAAAATTGGAGTGGAGCGCCATCCTGC
AGAAGATGGCGAGCGACCTTGGATTCTCGAAGATCCTGTTCGGCCTGTTGCCTAAGGACAGCCAGGACTA
CGAGAACGCCTTCATCGTCGGCAACTACCCGGCCGCCTGGCGCGAGCATTACGACCGGGCTGGCTACGCG
CGGGTCGACCCGACGGTCAGTCACTGTACCCAGAGCGTACTGCCGATTTTCTGGGAACCGTCCATCTACC
AGACGCGAAAGCAGCACGAGTTCTTCGAGGAAGCCTCGGCCGCCGGCCTGGTGTATGGGCTGACCATGCC
GCTGCATGGTGCTCGCGGCGAACTCGGCGCGCTGAGCCTCAGCGTGGAAGCGGAAAACCGGGCCGAGGCC
AACCGTTTCATGGAGTCGGTCCTGCCGACCCTGTGGATGCTCAAGGACTACGCACTGCAGAGCGGTGCCG
GACTGGCCTTCGAACATCCGGTCAGCAAACCGGTGGTTCTGACCAGCCGGGAGAAGGAAGTGTTGCAGTG
GTGCGCCATCGGCAAGACCAGTTGGGAGATATCGGTTATCTGCAACTGCTCGGAAGCCAATGTGAACTTC
CATATGGGAAATATTCGGCGGAAGTTCGGTGTGACCTCCCGCCGCGTAGCGGCCATTATGGCCGTTAATT
TGGGTCTTATTACTCTCTGA
>gi|110645304|ref|NC_002516.2|:c1559122-1558880 Pseudomonas aeruginosa PAO1 chromosome, complete genome
ATGGCTTCACACGAGAGAACACAGCCCCAAAACATGGCCTTCCGGGCAAAGGCCACCCGCACCGCCCGAC
GGGAAAGCCAGGAAACTTTCTGGAGCCGCTTCGGGATAAGCCAATCCTGCGGCAGTCGTTTCGAGAATGG
CGAGAACCTGCCCTTCCCTATATATCTGCTTTTGCATTTCTATATAGAAGGGCAAATTACCGATCGCCAG
CTCGCCGACCTGAGAGGCAAGATCAGAGAGTAA
>gi|110645304|ref|NC_002516.2|:1559254-1559859 Pseudomonas aeruginosa PAO1 chromosome, complete genome
ATGATCGTACAAATTGGTCGGCGCGAAGAGTTCGATAAAAAACTGCTGGGCGAGATGCACAAGTTGCGTG
CTCAAGTGTTCAAGGAGCGCAAAGGCTGGGACGTTAGTGTCATCGACGAGATGGAAATCGATGGTTATGA
CGCACTCAGTCCTTATTACATGTTGATCCAGGAAGATACTCCTGAAGCCCAGGTTTTCGGTTGCTGGCGA
ATTCTCGATACCACTGGCCCCTACATGCTGAAGAACACCTTCCCGGAGCTTCTGCACGGCAAGGAAGCGC
CTTGCTCGCCGCACATCTGGGAACTCAGCCGTTTCGCCATCAACTCTGGACAGAAAGGCTCGCTGGGCTT
TTCCGACTGTACGCTGGAGGCGATGCGCGCGCTGGCCCGCTACAGCCTGCAGAACGACATCCAGACGCTG
GTGACGGTAACCACCGTAGGCGTGGAGAAGATGATGATCCGTGCCGGCCTGGACGTATCGCGCTTCGGTC
CGCACCTGAAGATCGGCATCGAGCGCGCGGTGGCCTTGCGCATCGAACTCAATGCCAAGACCCAGATCGC
GCTTTACGGGGGAGTGCTGGTGGAACAGCGACTGGCGGTTTCATGA
>gi|110645304|ref|NC_002516.2|:c1561918-1559966 Pseudomonas aeruginosa PAO1 chromosome, complete genome
ATGTCACTGCTCAAGCAATTGTTCCTCGCCATCTGCCTGTTCCTCGTCGTCGCCTTCAGCGGCAGCTTCG
TCAGCAGCGTGGAGAACTCCCGCGAGCAGTTGCGCGGCCAGTTGCGCTCCCACGCCCAGGACGCCGCCAC
CGCCCTCGGCCTGTCCCTGACGCCACACGTGGACGACCCGGCGATGGTGCAACTGATGGTCAGCTCGATC
TTCGACAGCGGCTACTTCGCCAGCATCCGGGTGATCGACATCAAGAGCGGCAAGCCGCTGGTCGAGCGCG
TGCAGGCGCATGCCGAGCGTACGGTGCCCGGCTGGTTCGAACGCCTGGTGGATCTCCAGCCCCAGGGTGG
CGACGCGCTGATCATGCGCGGCTGGGAACAGGCGGCGCGGGTCGAGGTGGTCAGCCATCCGCAATTCGCC
CTGGCGCGCCTGTGGGACAGCGCCCTGGGCAGCCTCTACTGGTTGCTGGCCTGCGGCGCGGCGAGCCTGC
TGCTTGGCGGCTGGCTGCTGCGCCGCCAACTGCGCCCGCTCGACCAGATGGTGCGCCAGGCCCATGCCAT
CAGCCGGCGGGAATTCCTCAGCCTGCCCAGGCTGCCGCGCACGCCGGAGCTGCGCCGCGTGGTGCAGGCG
ATGAACCAGATGGTGGAGAAACTTCGCACGCTGTTCGCCGAGGAGGCTGCGCGTAGCGACAAGCTACGCG
CCCAGGCCTACCAGGACAGTCTCACCGGCCTGCCCAACCGGCGCCTGTTCGACGCCCGCCTCAACGAACA
GCTGGGTGCCGGCGAGCACGAGCATGCCGGGCAGCTCCTGCTGCTGCGCCTGAACGACCTGAACGGGCTG
AACCAGCGTCTCGGCGGGCAACGCACGGATGAATTGATCCAGGCGGTCGCCCGCTTGCTGGTCGACTCCT
GCGGCCAGCAAGGGCGCGCCGACTGGCTGCTGGCCCGCAGCCGCGGCGGCGAGTTCGCCGTGCTGGCTCC
CGGTTGCTCCCGCGAGCAGGCCGAACGCCTGGCCGAGGAACTCTGCGAAGGCCTGGAGAACCTTGCCCGC
ACCGGTGCCAGCGATCTCACGCCGGTCGCCTATCTCGGTATCAGCGCGTTCGCCGAAGGCGACTCGCCCG
CGGATCTCCTGGCCCGTGCCGACCAGGCCCTGGCCCAGGCCGAAAGCCAGCCCGCCCAGCCCTGGGCGAG
```

A Pseudomonas aeruginosa nevű baktérium NC_002516 nevű fnn fájl részlete. Jól láthatóak a FASTA formátumra jellemző, '>' jellel bevezetett fejléc sorok, mely ebben az esetben a forrás baktérium nevét, azonosítóját és a részszekvenciák helyét tartalmazza. A pirossal kiemelt rész egy LasR-RsaL-LasI fehérje hármashoz tartozó DNS szekvencia.

DOI:10.15774/PPKE.ITK.2014.006

IV. Melléklet

```
LOCUS       NC_002516            6264404 bp    DNA     circular BCT 22-DEC-2012
DEFINITION  Pseudomonas aeruginosa PAO1 chromosome, complete genome.
ACCESSION   NC_002516
VERSION     NC_002516.2  GI:110645304
DBLINK      Project: 57945
            BioProject: PRJNA57945
KEYWORDS    .
SOURCE      Pseudomonas aeruginosa PAO1
  ORGANISM  Pseudomonas aeruginosa PAO1
            Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales;
            Pseudomonadaceae; Pseudomonas.
REFERENCE   1  (bases 1 to 6264404)
  AUTHORS   Winsor,G.L., Van Rossum,T., Lo,R., Khaira,B., Whiteside,M.D.,
            Hancock,R.E. and Brinkman,F.S.
  TITLE     Pseudomonas Genome Database: facilitating user-friendly,
            comprehensive comparisons of microbial genomes
  JOURNAL   Nucleic acids Res. 37 (DATABASE ISSUE), D483-D488 (2009)
   PUBMED   18978025
REFERENCE   2  (bases 1 to 6264404)
  AUTHORS   Cirz,R.T., O'Neill,B.M., Hammond,J.A., Head,S.R. and Romesberg,F.E.
  TITLE     Defining the Pseudomonas aeruginosa SOS response and its role in
            the global response to the antibiotic ciprofloxacin
  JOURNAL   J. Bacteriol. 188 (20), 7101-7110 (2006)
   PUBMED   17015649
```

```
   gene            1559254..1559859
                   /gene="lasI"
                   /locus_tag="PA1432"
                   /db_xref="GeneID:881777"
   CDS             1559254..1559859
                   /gene="lasI"
                   /locus_tag="PA1432"
                   /note="Product name confidence: class 1 (Function
                   experimentally demonstrated in P. aeruginosa)"
                   /codon_start=1
                   /transl_table=11
                   /product="autoinducer synthesis protein LasI"
                   /protein_id="NP_250123.1"
                   /db_xref="GI:15596629"
                   /db_xref="GeneID:881777"
                   /translation="MIVQIGRREEFDKKLLGEMHKLRAQVFKERKGWDVSVIDEMEID
                   GYDALSPYYMLIQEDTPEAQVFGCWRILDTTGPYMLKNTFPELLHGKEAPCSPHIWEL
                   SRFAINSGQKGSLGFSDCTLEAMRALARYSLQNDIQTLVTVTTVGVEKMMIRAGLDVS
                   RFGPHLKIGIERAVALRIELNAKTQIALYGGVLVEQRLAVS"
   misc_feature    1559254..1559829
                   /gene="lasI"
                   /locus_tag="PA1432"
                   /note="N-acyl-L-homoserine lactone synthetase [Signal
                   transduction mechanisms / Secondary metabolites
                   biosynthesis, transport, and catabolism]; Region: LasI;
                   COG3916"
                   /db_xref="CDD:33702"
```

A Pseudomonas aeruginosa nevű baktérium NC_002516 nevű gbk fájl két részlete. Az elsőn a fájl eleje látható, ahol a GenBank rekord adatai és a forrás élőlény tulajdonságai után a kapcsolódó publikációk listája található. A második részleten pedig a LasI fehérjét kódoló génről tárolt információk láthatóak.

DOI:10.15774/PPKE.ITK.2014.006

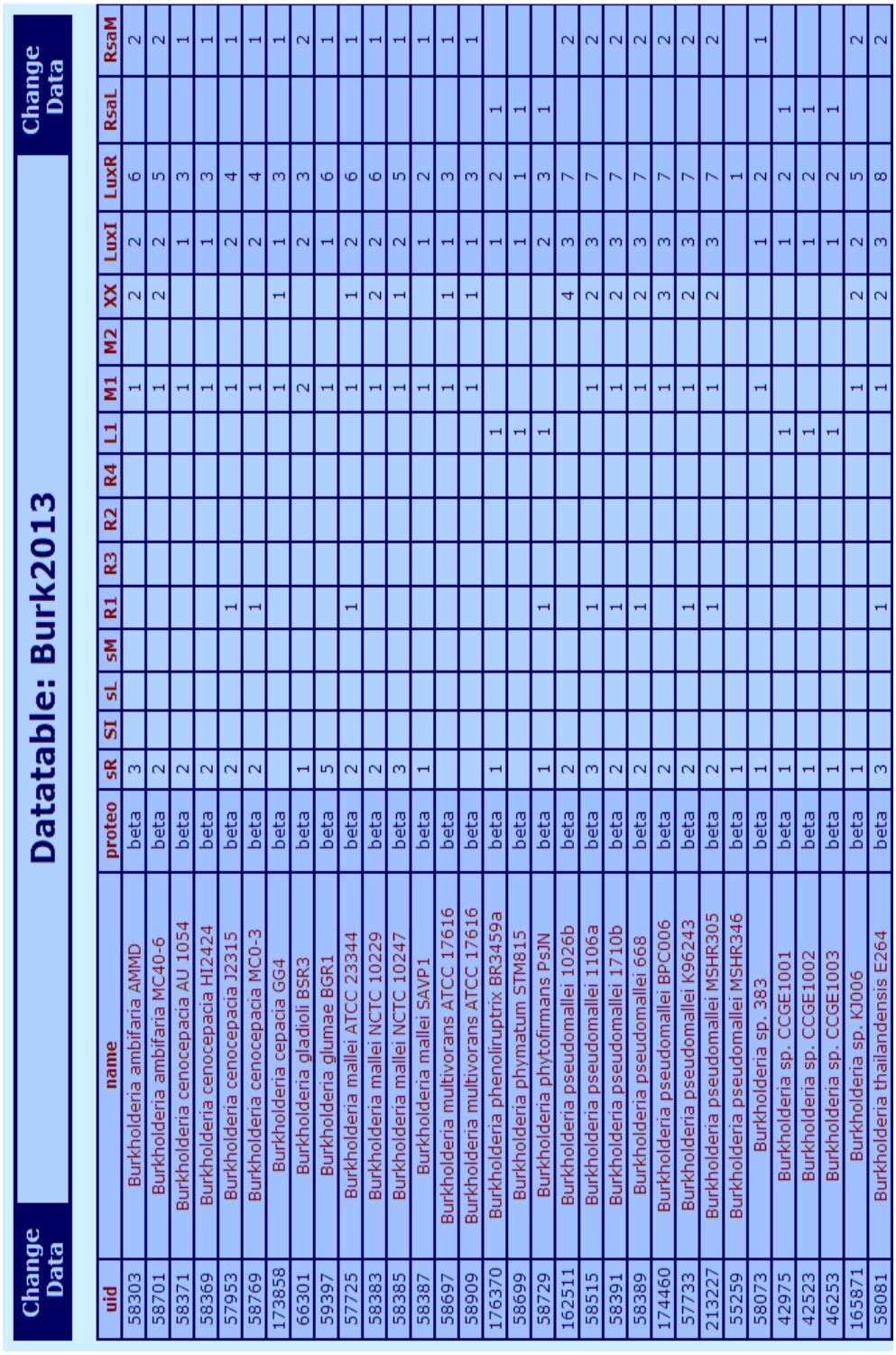
V. Melléklet

1536457..1537650	-	397	15596609	-	PA1412	-	COG2814G	hypothetical protein
1537755..1538627	+	290	15596610	-	PA1413	-	COG0583K	transcriptional regulator
1538683..1538916	+	77	15596611	-	PA1414	-	-	hypothetical protein
1538976..1539704	-	242	15596612	-	PA1415	-	COG0491R	hypothetical protein
1539749..1541131	-	460	15596613	-	PA1416	-	COG0277C	hypothetical protein
1541145..1542746	-	533	15596614	-	PA1417	-	COG0028EH	hypothetical protein
1542822..1544213	-	463	15596615	-	PA1418	-	COG0591ER	sodium:solute symport protein
1544307..1545818	-	503	15596616	-	PA1419	-	COG1457F	transporter
1545837..1546256	-	139	15596617	-	PA1420	-	COG4319S	hypothetical protein
1546271..1547230	-	319	15596618	gbuA	PA1421	-	COG0010E	guanidinobutyrase
1547437..1548330	+	297	15596619	gbuR	PA1422	-	COG0583K	protein GbuR
1548334..1549587	-	417	15596620	bdlA	PA1423	-	COG0840NT	BdlA
1549883..1550494	+	203	15596621	-	PA1424	-	COG3148S	hypothetical protein
1550984..1552600	+	538	15596622	-	PA1425	-	COG0488R	ABC transporter ATP-binding protein
1552641..1552997	+	118	15596623	-	PA1426	-	-	hypothetical protein
1553112..1553675	+	187	15596624	-	PA1427	-	-	hypothetical protein
1554415..1554846	+	143	15596625	-	PA1428	-	COG0456R	hypothetical protein
1555012..1557720	+	902	15596626	-	PA1429	-	COG0474P	cation-transporting P-type ATPase
1558171..1558890	+	239	15596627	lasR	PA1430	-	COG2771K	transcriptional regulator LasR
1558880..1559122	-	80	15596628	rsaL	PA1431	-	-	regulatory protein RsaL
1559254..1559859	+	201	15596629	lasI	PA1432	-	COG3916TQ	autoinducer synthesis protein LasI
1559966..1561918	-	650	15596630	-	PA1433	-	COG2200T	hypothetical protein
1561919..1562632	-	237	15596631	-	PA1434	-	COG3672S	hypothetical protein
1562813..1563970	+	385	15596632	-	PA1435	-	COG0845M	Resistance-Nodulation-cell Division (RND)
1563967..1567077	+	1036	15596633	-	PA1436	-	COG0841V	Resistance-Nodulation-cell Division (RND)
1567181..1567870	+	229	15596634	-	PA1437	-	COG0745TK	two-component response regulator
1567848..1569293	+	481	15596635	-	PA1438	-	COG0642T	two-component sensor
1569302..1569709	-	135	15596636	-	PA1439	-	COG2832S	hypothetical protein
1569721..1570308	-	195	15596637	-	PA1440	-	COG3318R	hypothetical protein
1570496..1571779	+	427	15596638	-	PA1441	-	COG3144N	hypothetical protein
1572023..1572544	+	173	15596639	-	PA1442	-	COG1580N	flagellar basal body protein FliL
1572552..1573523	+	323	15596640	fliM	PA1443	-	COG1868N	flagellar motor switch protein FliM
1573551..1574024	+	157	15596641	fliN	PA1444	-	COG1886NU	flagellar motor switch protein
1574026..1574478	+	150	15596642	fliO	PA1445	-	COG3190N	flagellar protein FliO
1574475..1575242	+	255	15596643	fliP	PA1446	-	COG1338NU	flagellar biosynthesis protein FliP
1575290..1575559	+	89	15596644	fliQ	PA1447	-	COG1987NU	flagellar biosynthesis protein FliQ
1575559..1576335	+	258	15596645	fliR	PA1448	-	COG1684NU	flagellar biosynthesis protein FliR
1576338..1577474	+	378	15596646	flhB	PA1449	-	COG1377NU	flagellar biosynthesis protein FlhB
1577547..1578806	+	419	15596647	-	PA1450	-	COG1512R	hypothetical protein
1578839..1580182	+	447	15596648	-	PA1451	-	COG1512R	hypothetical protein
1580321..1582444	+	707	15596649	flhA	PA1452	-	COG1298NU	flagellar biosynthesis protein FlhA
1582528..1583817	+	429	15596650	flhF	PA1453	-	COG1419N	flagellar biosynthesis regulator FlhF
1583956..1584798	+	280	15596651	fleN	PA1454	-	COG0455D	flagellar synthesis regulator FleN
1584795..1585538	+	247	15596652	fliA	PA1455	-	COG1191K	flagellar biosynthesis sigma factor
1585640..1586014	+	124	15596653	cheY	PA1456	-	COG2204T	chemotaxis protein CheY
1586034..1586822	+	262	15596654	cheZ	PA1457	-	COG3143NT	chemotaxis protein CheZ
1587023..1589284	+	753	15596655	-	PA1458	-	COG0643NT	two-component sensor
1589338..1590444	+	368	15596656	-	PA1459	-	COG2201NT	chemotaxis-specific methylesterase
1590533..1591273	+	246	15596657	motC	PA1460	-	COG1291N	flagellar motor protein
1591286..1592176	+	296	15596658	motD	PA1461	-	COG1360N	flagellar motor protein MotD
1592271..1593059	+	262	15596659	-	PA1462	-	COG1192D	plasmid partitioning protein
1593151..1594041	+	296	15596660	-	PA1463	-	COG0835NT	hypothetical protein
1594087..1594566	+	159	15596661	-	PA1464	-	COG0835NT	purine-binding chemotaxis protein
1594597..1595004	+	135	15596662	-	PA1465	-	-	hypothetical protein
1595032..1595718	-	228	15596663	-	PA1466	-	COG0625O	hypothetical protein
1595827..1596798	+	323	15596664	-	PA1467	-	COG2378K	hypothetical protein
1596889..1597305	+	138	15596665	-	PA1468	-	-	hypothetical protein
1597365..1598087	+	240	15596666	-	PA1469	-	COG1496S	hypothetical protein
1598221..1598958	+	245	15596667	-	PA1470	-	COG1028IQR	short-chain dehydrogenase
1599024..1599320	-	98	15596668	-	PA1471	-	-	hypothetical protein
1599428..1599982	-	184	15596669	-	PA1472	-	COG1670J	hypothetical protein

A Pseudomonas aeruginosa nevű baktérium NC_002516 nevű ptt fájl részlete. A táblázatos formában tárolt adatok mind kézi mind automatikus adatgyűjtés esetén is könnyen kigyűjthetőek. A pirossal kiemelt rész egy LasR-RsaL-LasI fehérje hármas.

DOI:10.15774/PPKE.ITK.2014.006

VI. Melléklet



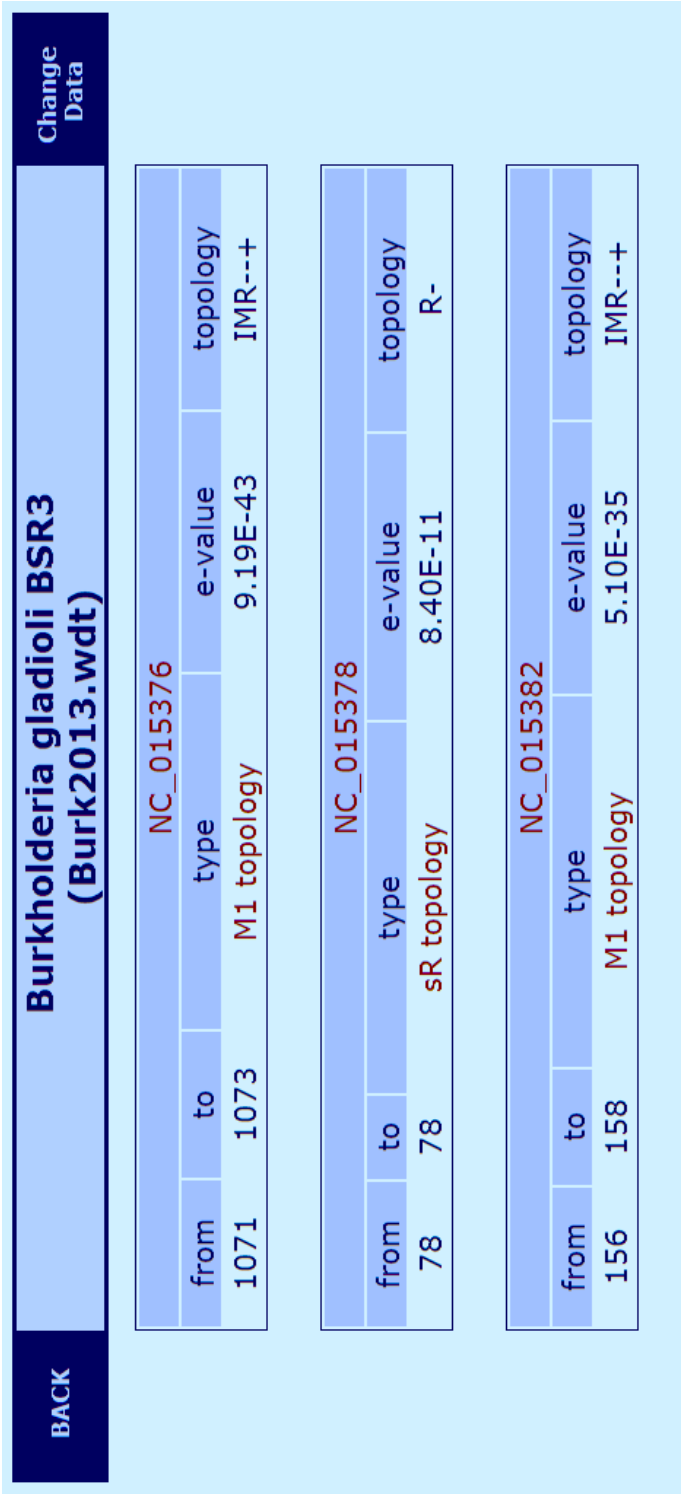
Change Data | Datatable: Burk2013 | Change Data

uid	name	proteo	sR	SI	sL	sM	R1	R3	R2	R4	L1	M1	M2	XX	LuxI	LuxR	RsaL	RsaM
58303	Burkholderia ambifaria AMMD	beta	3									1		2	2	6		2
58701	Burkholderia ambifaria MC40-6	beta	2									1		2	2	5		2
58371	Burkholderia cenocepacia AU 1054	beta	2									1			1	3		1
58369	Burkholderia cenocepacia HI2424	beta	2									1			1	3		1
57953	Burkholderia cenocepacia J2315	beta	2				1					1			2	4		1
58769	Burkholderia cenocepacia MC0-3	beta	2				1					1			2	4		1
173858	Burkholderia cepacia GG4	beta										1		1	1	3		1
66301	Burkholderia gladioli BSR3	beta	1									2			2	3		2
59397	Burkholderia glumae BGR1	beta	5									1			1	6		1
57725	Burkholderia mallei ATCC 23344	beta	2				1					1		1	2	6		1
58383	Burkholderia mallei NCTC 10229	beta	2									1		2	2	6		1
58385	Burkholderia mallei NCTC 10247	beta	3									1		1	2	5		1
58387	Burkholderia mallei SAVP1	beta	1									1			1	2		1
58697	Burkholderia multivorans ATCC 17616	beta										1		1	1	3		1
58909	Burkholderia multivorans ATCC 17616	beta										1		1	1	3		1
176370	Burkholderia phenoliruptrix BR3459a	beta	1								1				1	2	1	
58699	Burkholderia phymatum STM815	beta									1				1	1	1	
58729	Burkholderia phytofirmans PsJN	beta	1				1				1				2	3	1	
162511	Burkholderia pseudomallei 1026b	beta	2											4	3	7		2
58515	Burkholderia pseudomallei 1106a	beta	3				1					1		2	3	7		2
58391	Burkholderia pseudomallei 1710b	beta	2				1					1		2	3	7		2
58389	Burkholderia pseudomallei 668	beta	2				1					1		2	3	7		2
174460	Burkholderia pseudomallei BPC006	beta	2									1		3	3	7		2
57733	Burkholderia pseudomallei K96243	beta	2				1					1		2	3	7		2
213227	Burkholderia pseudomallei MSHR305	beta	2				1					1		2	3	7		2
55259	Burkholderia pseudomallei MSHR346	beta	1													1		
58073	Burkholderia sp. 383	beta	1									1			1	2		1
42975	Burkholderia sp. CCGE1001	beta	1								1				1	2	1	
42523	Burkholderia sp. CCGE1002	beta	1								1				1	2	1	
46253	Burkholderia sp. CCGE1003	beta	1								1				1	2	1	
165871	Burkholderia sp. KJ006	beta	1									1		2	2	5		2
58081	Burkholderia thailandensis E264	beta	3				1					1		2	3	8		2

A vizualizációs honlap főtáblázatának részlete. Egy burkholdéria baktérium törzsön való keresés eredményét láthatjuk. A piros szín jelzi az adott szöveg link mivoltát.

DOI:10.15774/PPKE.ITK.2014.006

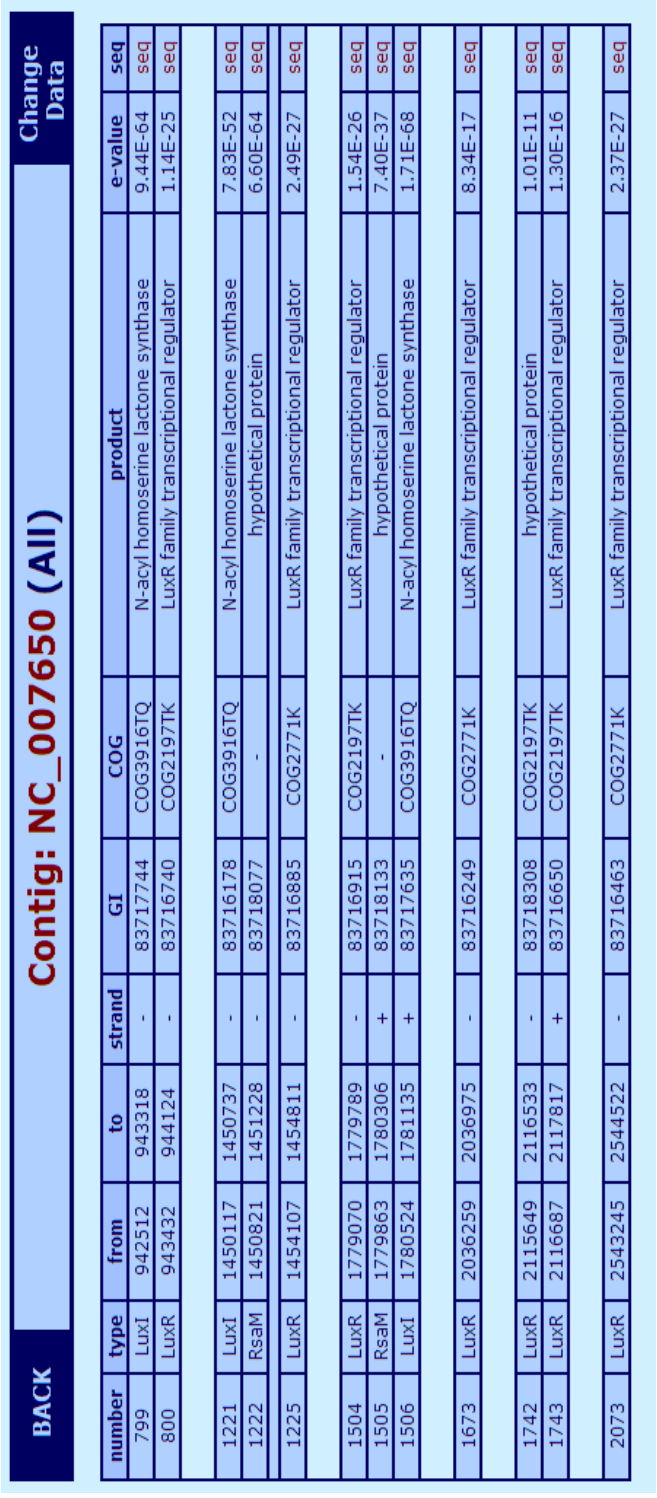
VII. Melléklet



A vizualizációs honlap egy baktériumról szóló oldala (jelen esetben a Burkholderia gladioli BSR3). Jól látható, hogy 3 különböző contigban is találtunk quorum sensing gént.

DOI:10.15774/PPKE.ITK.2014.006

VIII. Melléklet



BACK

Contig: NC_007650 (All)

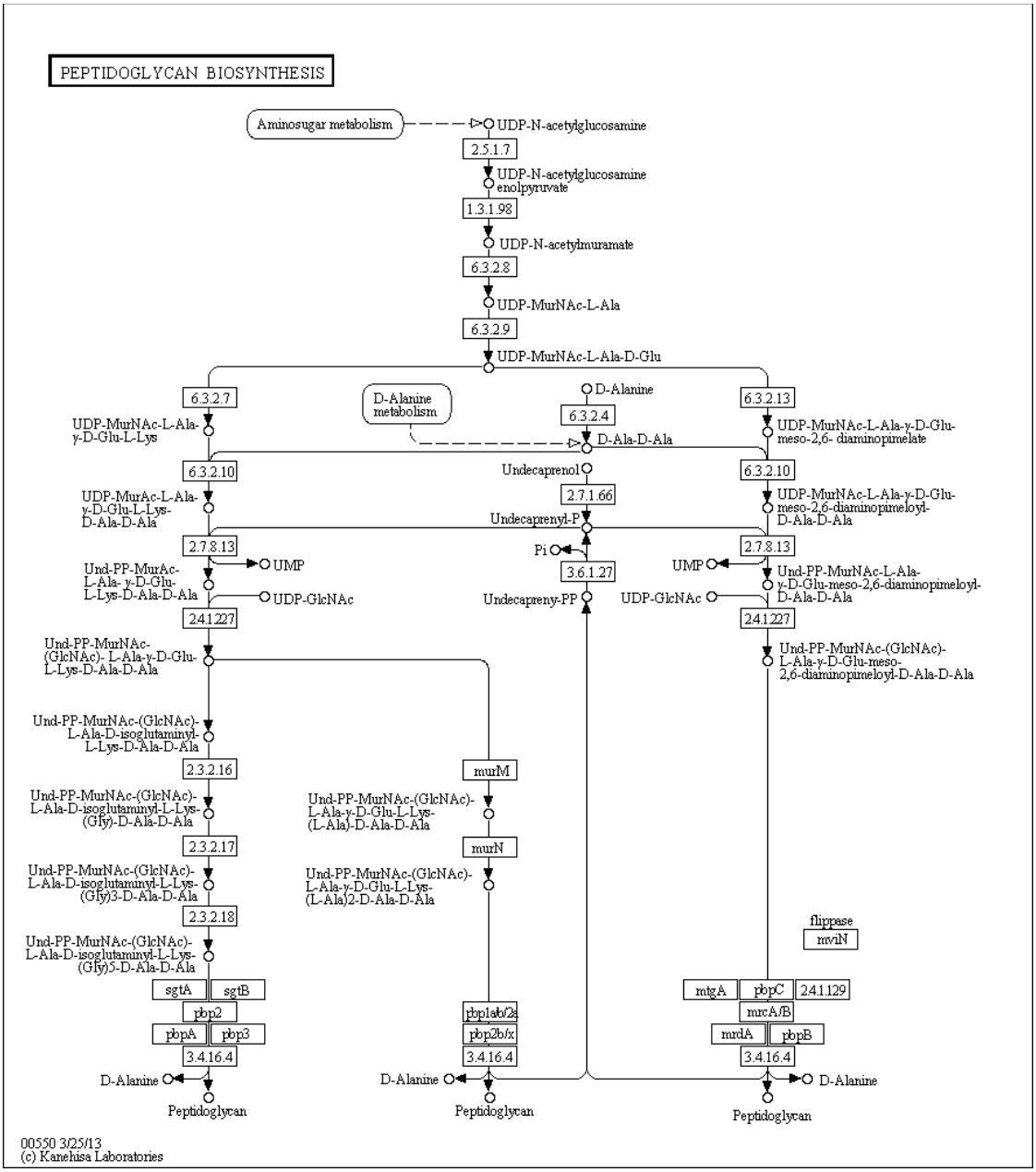
Change Data

number	type	from	to	strand	GI	COG	product	e-value	seq
799	LuxI	942512	943318	-	83717744	COG3916TQ	N-acyl homoserine lactone synthase	9.44E-64	seq
800	LuxR	943432	944124	-	83716740	COG2197TK	LuxR family transcriptional regulator	1.14E-25	seq
1221	LuxI	1450117	1450737	-	83716178	COG3916TQ	N-acyl homoserine lactone synthase	7.83E-52	seq
1222	RsaM	1450821	1451228	-	83718077	-	hypothetical protein	6.60E-64	seq
1225	LuxR	1454107	1454811	-	83716885	COG2771K	LuxR family transcriptional regulator	2.49E-27	seq
1504	LuxR	1779070	1779789	-	83716915	COG2197TK	LuxR family transcriptional regulator	1.54E-26	seq
1505	RsaM	1779863	1780306	+	83718133	-	hypothetical protein	7.40E-37	seq
1506	LuxI	1780524	1781135	+	83717635	COG3916TQ	N-acyl homoserine lactone synthase	1.71E-68	seq
1673	LuxR	2036259	2036975	-	83716249	COG2771K	LuxR family transcriptional regulator	8.34E-17	seq
1742	LuxR	2115649	2116533	-	83718308	COG2197TK	hypothetical protein	1.01E-11	seq
1743	LuxR	2116687	2117817	+	83716650	COG2197TK	LuxR family transcriptional regulator	1.30E-16	seq
2073	LuxR	2543245	2544522	-	83716463	COG2771K	LuxR family transcriptional regulator	2.37E-27	seq

A vizualizációs honlap egy contigról szóló oldala. Megfigyelhető a táblázaton a topológiák közötti nagyobb távolság jelölése (világoskék) és a topológián belüli kisebb hézag jelölése (sötétkék)

DOI:10.15774/PPKE.ITK.2014.006

IX. Melléklet



A baktériumok peptidoglycan bioszintézisének metabolikus útvonal térképe.

DOI:10.15774/PPKE.ITK.2014.006

X. Melléklet

A HMM profilok szekvencia azonosítóinak (GI szám) listája

LuxR fehérje család

Burkholderia ambifaria AMMD	115360781
Burkholderia CCGE1001	323528655
Citrobacter koseri ATCC BAA 895	157145292
Citrobacter rodentium ICC168	283785661
Cronobacter sakazakii ATCC BAA 894	156933489
Escherichia coli	253773131
Escherichia fergusonii ATCC 35469	218548547
Listonella anguillarum M3	537456902
Nitrosospira multiformis ATCC 25196	82703532
Pseudomonas aeruginosa B136 33	478479802
Pseudomonas aeruginosa PAO1	15596627
Ralstonia solanacearum FQY 4	525711912
Ralstonia solanacearum GMI1000	17548999
Ralstonia solanacearum PSI07	300693708
Rhodobacter sphaeroides ATCC 17029	126462526
Shigella boydii CDC 3083 94	187730097
Shigella sonnei Ss046	74311738
Thiomonas 3As	410693608
Vibrio anguillarum 775	336125724

RsaM fehérje család

Burkholderia ambifaria AMMD	115360795
Burkholderia pseudomallei 1106a	126456722
Burkholderia pseudomallei 1710b	76819132
Burkholderia pseudomallei K96243	53722207
Pseudomonas fuscovaginae	290454886

RsaL fehérje család

Burkholderia Kururiensis	172152204
Burkholderia Xenovorans LB400	91779484
Gallionella capsiferriformans ES/2	302878669
Pseudomonas aeruginosa PAO1	15596628
Pseudomonas fuscovaginae UPB0736	270341117
Pseudomonas putida	73672743

LuxI fehérje család

1. csoport

Acidovorax citrulli AAC00 1	120612452
Agrobacterium vitis S4	222109224
Agrobacterium vitis S4	222148403
Asticcacaulis excentricus CB 48	315499427
Beijerinckia indica ATCC 9039	182677911

DOI:10.15774/PPKE.ITK.2014.006

Bradyrhizobium BTAi1	148258316
Bradyrhizobium japonicum USDA 110	27376174
Bradyrhizobium ORS278	146338046
Burkholderia cenocepacia J2315	206562109
Burkholderia cenocepacia MC0 3	170736658
Burkholderia mallei ATCC 23344	53716259
Burkholderia phytofirmans PsJN	187922147
Burkholderia pseudomallei 1106a	126456405
Burkholderia pseudomallei 1710b	76819605
Burkholderia pseudomallei 668	126442432
Burkholderia pseudomallei K96243	53722591
Burkholderia thailandensis E264	83717744
Burkholderia xenovorans LB400	91780462
Chelativorans BNC1	110634657
Dinoroseobacter shibae DFL 12	159042868
Dinoroseobacter shibae DFL 12	159045391
Erwinia amylovora ATCC 49946	292899117
Erwinia amylovora CFBP1430	292487895
Erwinia billingiae Eb661	300716151
Erwinia pyrifoliae Ep1 96	259908859
Erwinia tasmaniensis Et1 99	188534208
Jannaschia CCS1	89053111
Ketogulonicigenium vulgare Y25	310816982
Mesorhizobium ciceri biovar biserrulae WSM1271	319785240
Mesorhizobium ciceri biovar biserrulae WSM1271	319785575
Mesorhizobium loti MAFF303099	13474693
Mesorhizobium loti MAFF303099	13475097
Mesorhizobium loti MAFF303099	13488405
Methylobacterium 4 46	170743556
Methylobacterium chloromethanicum CM4	218532853
Methylobacterium extorquens AM1	240141365
Methylobacterium extorquens AM1	240142369
Methylobacterium extorquens DM4	254563871
Methylobacterium extorquens PA1	163853909
Methylobacterium nodulans ORS 2060	220920329
Methylobacterium nodulans ORS 2060	220921757
Methylobacterium populi BJ001	188584214
Methylobacterium radiotolerans JCM 2831	170745592
Methylobacterium radiotolerans JCM 2831	170752130
Methylobacterium radiotolerans JCM 2831	170752149
Methylocella silvestris BL2	217976295
Nitrobacter winogradskyi Nb 255	75674824
Pantoea At 9b	317047631
Pantoea vagans C9 1	308186329
Paracoccus denitrificans PD1222	119383539
Phenylobacterium zucineum HLK1	197103146
Pseudomonas aeruginosa LESB58	218890276
Pseudomonas aeruginosa PA7	152983670

DOI:10.15774/PPKE.ITK.2014.006

Pseudomonas aeruginosa PAO1	15598672
Pseudomonas aeruginosa UCBPP PA14	116051495
Ralstonia solanacearum GMI1000	17549000
Ralstonia solanacearum PSI07	300693709
Rhizobium etli CFN 42	86358519
Rhizobium etli CFN 42	86361170
Rhizobium etli CIAT 652	190892658
Rhizobium etli CIAT 652	190894997
Rhizobium leguminosarum bv trifolii WSM1325	241205632
Rhizobium leguminosarum bv trifolii WSM2304	209550248
Rhizobium leguminosarum bv viciae 3841	116253120
Rhizobium NGR234	227822237
Rhodobacter capsulatus SB 1003	294675886
Rhodomicrobium vannielii ATCC 17100	312115397
Rhodopseudomonas palustris BisA53	115524333
Rhodopseudomonas palustris BisB18	90422222
Rhodopseudomonas palustris BisB5	91974883
Rhodopseudomonas palustris BisB5	91976652
Rhodopseudomonas palustris CGA009	39933397
Rhodopseudomonas palustris DX 1	316931699
Rhodopseudomonas palustris DX 1	316931972
Rhodopseudomonas palustris DX 1	316933443
Rhodopseudomonas palustris HaA2	86747543
Rhodopseudomonas palustris HaA2	86750431
Rhodopseudomonas palustris TIE 1	192288753
Rhodopseudomonas palustris TIE 1	192290719
Rhodospirillum rubrum ATCC 11170	83594725
Roseobacter denitrificans OCh 114	110678945
Ruegeria pomeroyi DSS 3	56695287
Ruegeria pomeroyi DSS 3	56697148
Sinorhizobium medicae WSM419	150396770
Sinorhizobium meliloti 1021	15965592
Sphingobium japonicum UT26S	294012985
Sphingopyxis alaskensis RB2256	103488005
Sphingopyxis alaskensis RB2256	103488067

2. csoport

Aeromonas hydrophila ATCC 7966	117620619
Aeromonas salmonicida A449	145300630
Aliivibrio salmonicida LFI1238	209809707
Bradyrhizobium BTAi1	148241068
Ralstonia solanacearum CFBP2957	300702553
Ralstonia solanacearum GMI1000	17548003
Ralstonia solanacearum PSI07	300689865
Shewanella violacea DSS12	294142231
Shewanella woodyi ATCC 51908	170727338
Vibrio fischeri ES114	59714107
Vibrio fischeri MJ11	197337622

DOI:10.15774/PPKE.ITK.2014.006

3. csoport

Agrobacterium radiobacter K84	222081959
Candidatus Hamiltonella defensa 5AT	238898059
Chromobacterium violaceum ATCC 12472	34499546
Citrobacter rodentium ICC168	283786605
Desulfovibrio magneticus RS 1	239908484
Dickeya dadantii 3937	307133125
Dickeya dadantii Ech586	271498690
Dickeya zeae Ech1591	251787752
Edwardsiella ictaluri 93 146	238920761
Edwardsiella tarda EIB202	269139939
Enterobacter cloacae SCF1	311279343
Erwinia billingiae Eb661	300715700
Erwinia tasmaniensis Et1 99	188533100
Geobacter FRC 32	222055509
Geobacter uraniireducens Rf4	148265143
Pantoea ananatis LMG 20103	291617508
Pantoea vagans C9 1	298717241
Pectobacterium atrosepticum SCRI1043	50119066
Pectobacterium carotovorum PC1	253690508
Pectobacterium wasabiae WPP163	261823622
Pseudomonas syringae B728a	66044866
Pseudomonas syringae phaseolicola 1448A	71733512
Pseudomonas syringae tomato DC3000	28871017
Rhodopseudomonas palustris BisB18	90423533
Serratia proteamaculans 568	157368316
Sodalis glossinidius morsitans	85058262
Variovorax paradoxus S110	239820478
Yersinia enterocolitica 8081	123441909
Yersinia pestis Angola	162418121
Yersinia pestis Angola	162420736
Yersinia pestis Antiqua	108806278
Yersinia pestis Antiqua	108807949
Yersinia pestis biovar Microtus 91001	45442065
Yersinia pestis biovar Microtus 91001	45443198
Yersinia pestis CO92	218928156
Yersinia pestis CO92	218929545
Yersinia pestis KIM 10	22125627
Yersinia pestis KIM 10	22127250
Yersinia pestis Nepal516	108812214
Yersinia pestis Nepal516	108813342
Yersinia pestis Pestoides F	145597842
Yersinia pestis Pestoides F	145599153
Yersinia pestis Z176003	294502864
Yersinia pestis Z176003	294504216
Yersinia pseudotuberculosis IP 31758	153947446
Yersinia pseudotuberculosis IP 31758	153949171
Yersinia pseudotuberculosis IP 32953	51596820

DOI:10.15774/PPKE.ITK.2014.006

Yersinia pseudotuberculosis IP 32953	51597560
Yersinia pseudotuberculosis PB1	186895897
Yersinia pseudotuberculosis PB1	186896694
Yersinia pseudotuberculosis YPIII	170023041
Yersinia pseudotuberculosis YPIII	170023897

4. csoport

Acidithiobacillus ferrooxidans ATCC 23270	218667155
Acidithiobacillus ferrooxidans ATCC 53993	198283774
Acinetobacter baumannii AB0057	213155520
Acinetobacter baumannii AB307 0294	215485040
Acinetobacter baumannii ACICU	184156456
Acinetobacter baumannii ATCC 17978	126640214
Acinetobacter baumannii AYE	169797686
Acinetobacter DR1	299771986
Burkholderia 383	78061900
Burkholderia ambifaria AMMD	115358867
Burkholderia ambifaria MC40 6	172063604
Burkholderia CCGE1001	323528657
Burkholderia CCGE1002	295700195
Burkholderia CCGE1003	307727056
Burkholderia cenocepacia AU 1054	107025995
Burkholderia cenocepacia HI2424	116692820
Burkholderia cenocepacia J2315	206563718
Burkholderia cenocepacia MC0 3	170737929
Burkholderia glumae BGR1	238024811
Burkholderia mallei ATCC 23344	53717162
Burkholderia mallei NCTC 10229	124381232
Burkholderia mallei NCTC 10247	126446861
Burkholderia mallei SAVP1	121598144
Burkholderia multivorans ATCC 17616	161520517
Burkholderia multivorans ATCC 17616	189353293
Burkholderia phymatum STM815	186473255
Burkholderia phytofirmans PsJN	187918991
Burkholderia pseudomallei 1106a	126457870
Burkholderia pseudomallei 1710b	76818733
Burkholderia pseudomallei 668	126444167
Burkholderia pseudomallei K96243	53721909
Burkholderia thailandensis E264	83717635
Burkholderia vietnamiensis G4	134293940
Burkholderia xenovorans LB400	91779485
Gallionella capsiferriformans ES 2	302878670
Halothiobacillus neapolitanus c2	261855567
Pseudomonas aeruginosa LESB58	218892696
Pseudomonas aeruginosa PA7	152986918
Pseudomonas aeruginosa PAO1	15596629
Pseudomonas aeruginosa UCBPP PA14	116049377