A bakteriális kommunikáció és kooperáció génjeinek elhelyezkedése ismert genomokban.

Az AHL szabályzórendszer génjei.

Doktori disszertáció tézisei

Pázmány Péter Katolikus Egyetem

Információs Technológiai és Bionikai Kar

Multidiszciplináris Műszaki és Természettudományi Doktori Iskola

Gelencsér Zsolt

Konzulens: Prof. Pongor Sándor

2014

1. Bevezetés

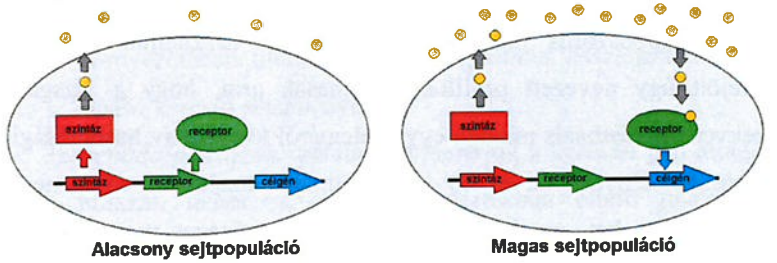
A szekvenálási módszerek közelmúltban történt hatalmas fejlődésének hatására a megismert DNS szekvenciák száma nagyon nagymértékben megnőtt: a tíz évvel ezelőtt rendelkezésre álló szekvenicák közel százszorosa a ma elérhetők száma, és az új módszerek lényegesen alacsonyabb költsége miatt ez a növekedés valószínűleg folytatódni fog. A DNS szekvenálásán kívül azonban lényeges még a szekvenciarészek pontos biológiai funkciójának a feltérképezése is. Ennek a folyamatnak egy fontos része számítógépes feladat: ezekkel az eszközökkel tudjuk a genomszekvenciát összevetni az ismert adatbázisokkal, és a csak részben automatizálható annotálási folyamatok során lépésenként felderíteni a genomban kódolt funkciókat.

A humán genommal kapcsolatos kutatások mellett a mai kutatás talán legfontosabb területe a baktériumok genomjának megismerése. Kiderült ugyanis, hogy a baktériumok nagyon bonyolult közösségeket alkotnak, - amelyek az emberi egészség fenntartásában, vagy éppen a fertőzések létrehozásában, sőt a talaj és a tengeri bioszféra egyensúlyában is szerepet játszanak - ehhez kommunikációra és kooperációra van szükség, amelynek molekuláris mechanizmusai nagy gyakorlati jelentőséggel bírnak.

A ma talán legjobban ismert kommunikációs mechanizmus az úgy nevezett *quorum sensing*, amelynek során a baktériumok külső kémiai jelanyagokkal érintkeznek egymással. Munkám során az

egyik legjobban ismert *quorum sensing* mechanizmus (az *N-acil homoserin laktón* jelanyagot használó úgy nevezett AHL rendszer) génjeit vizsgáltam, melyről éppen munkám során derült ki, hogy sok száz faj használja és igen nagy változatosságot mutat. Ennek a jelzőrendszernek a működése két fehérje családon alapul: a *LuxI* család tagjai szintetizálják a jelanyagot, ami egy kritikus koncentráció érték felett kapcsolatba lép a *LuxR* családba tartozó fehérjével: ez legtöbb esetben egy olyan komplexet eredményez, ami egy speciális DNS szakaszhoz kötődik a célgén promoter régiójában. Szabályozás szempontjából két fehérje között pozitív visszacsatolás van. Vannak baktériumok, amelyekben a *LuxR* és *LuxI* fehérjéken kívül egy harmadik, szorosan mellettük elhelyezkedő negatív szabályzó fehérje is megjelenik. Ezek közül két fontosat érdemes megemlíteni: az *RsaL* és az *RsaM*.

Munkám célja az volt, hogy az AHL rendszer génjeit megvizsgáljam a ma rendelkezésre álló összes bakteriális genomban, ezekhez megfelelő számítógépes programot tervezzek, és rendszerezzem a kapott adatokat.



A quorum sensing mechanizmusa

2. Módszerek

Munkám során jól ismert bioinformatikai alapalgoritmusokat használtam, melyeket program szkriptek segítségével kötöttem össze: az algoritmus kimenetéből a szükséges információkat kinyertem, és a következő algoritmus bementi formátumára hoztam. Ezzel a módszerrel egy kismértékű felhasználói beavatkozást igénylő munkafolyamatot hoztam létre.

A munkám során kifejlesztett genomannotálási algoritmus az alrendszerek azonosításán alapszik: nem egy genom vizsgálatát végezzük el, hanem egy meghatározott alrendszert vizsgálunk meg sok genomban. Az alrendszer egy génekből illetve azok kapcsolataiból álló szabálysorozat, amelyet kísérleti vizsgálatok alapján ismerünk meg, majd ezeket a géneket és kapcsolatokat keressük az eddig ismert genomok szekvenciáiban. Munkám során ezt az alrendszert a *quorum sensing* kommunikációért felelős fehérjecsaládok génjei alkották.

A keresés alapja az úgynevezett *rejtett Markov modell*, amelyet a bioinformatikában többek között az egy családhoz tartozó fehérjék többszörös illesztésének leírásához használjuk. Az így létrejött úgy nevezett profilok alkalmasak arra, hogy a vizsgált szekvencia-adatbázis minden egyes eleméről kvantitatív hasonlósági mérőszámot nyerjünk, és így kiválaszthatjuk azokat az eddig nem annotált géneket, amelyek szignifikáns egyezőséget mutatnak, és eleget tesznek az előzetesen felállított szabályoknak is.

3. Új tudományos eredmények

I. Tézis: Kidolgoztam egy számítógéppel automatizálható algoritmust, amely alkalmas egymással szomszédos gének és alrendszerek topológiájának elemzésére és képes a genomiális adatbázisok még nem annotált génjeinek felismerésére is.

Az algoritmus előfeltétele a fehérjecsaládok vagy az alrendszer génjeinek HMM profilja és a validációs tulajdonságok listája (például: maximális génhossz, maximális géntávolság ... stb.). A program során először a profilok alapján egy keresés történik az NCBI (National Center for Biotechnology Information) génadatbázisán. Ez a keresés viszonylag engedékeny, így a találatok egy validáláson mennek keresztül. Az ehhez szükséges adatok az NCBI különböző adatbázisából töltődnek le. Az ellenőrzés több módszerrel zajlik:

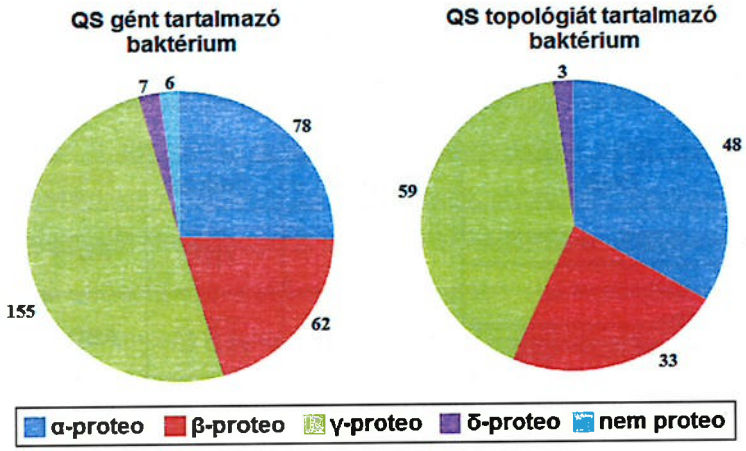
1) *Korábbi eredmények alapján*: például ha az adatbázis már tartalmazza az adott gén funkcióját, és ez a funkció hasonlóságot mutat az általunk keresettel.
2) *Más keresési algoritmusok segítségével:* például a találatok környezetében futatott BLAST algoritmus összehasonlítása a HMM keresés eredményeivel.
3) *Géntulajdonságok:* például ha ismerjük a keresett gén átlagos hosszát, akkor az ettől nagymértékben eltérő találatokat elvethetjük.

A validálás után a program az általam kifejlesztett algoritmus segítségével meghatározza a talált gének pontos topológiáját. Ezután egy jelentést készít az újonnan talált génekről és felismert topológia típusokról, mely egy webes környezetben könnyen áttekinthető.

II. Tézis: Megmutattam, hogy a jelenleg elérhető proteobaktérium genomok 12%-ában találhatóak AHL quorum sensing gének és ez összhangban van az erre vonatkozó biológiai becslésekkel, másrészt több tucat olyan feltételezhető AHL géncsoportot észleltem, melyek eddig nem voltak ismertek.

Az N-AHL alapú *quorum sensing* gének topológiai elrendeződésének elemzése a *Pseudomonasok* rendjébe tartozó baktériumok vizsgálatával indult, melynek keresési terét kiterjesztettem az összes elérhető teljes baktérium genomra. (A bakteriális genomok forrása az NCBI GenBank adatbázisa volt.) A folyamat során a *luxI* és a *luxR* gént is tartalmazó baktériumok mindegyike a proteobaktériumok törzsébe tartozott. Mivel létezik mind a *luxI*, mind a *luxR* gének által kódolt fehérjéhez nagyon hasonló, más funkcióval rendelkező fehérjecsalád, ezért a nem teljesen egyértelmű találati eredmények manuális ellenőrzését hossz és szekvencia-lefedettség ellenőrzésével végeztem. Az ellenőrzésekhez szigorú paramétereket állítottam be, hogy minél megbízhatóbb eredményhalmazt kapjak. A nem annotált gének közül csak azok kerülhettek be az eredmények közé, amelyek egy ismert elrendeződés részei voltak.

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 *luxI* és *luxR* gének egymás közelében helyezkednek el. Másodszor a keresés az ismert *LuxI* és *LuxR* fehérjékhez való hasonlóságon alapszik. Tehát kimaradtak azok a *luxR* gének, amelyek magányosan állnak vagy valószínű egy más típusú jeltermelést szabályoznak. 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 *strain*ek frekvenciájával (6-12%).



III. Tézis: Kidolgoztam egy jelölésrendszert, amely alkalmas a quorum sensing rendszerek és más kisméretű alrendszerek topológiájának felírására és szemléltetésére.

Az elrendeződéseket két szempont alapján vizsgáltam: a gének mennyire fedik át egymást illetve milyen az orientációjuk. (Ez az irányultság a gén kifejeződésének irányával van összefüggésben, ami attól függ, hogy melyik DNS szálon található.)

Munkám kezdetén érdemesnek láttam meghatározni egy egyszerűsített, formális felírást, amivel jelölni tudtam az adott elrendeződést. 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.

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

IV. Tézis: Megmutattam, hogy az egymás szomszédságában álló *luxI-luxR* párok illetve a hozzájuk csatlakozó szabályozógének (*rsaM* és *rsaL*) jellemző topológiai elrendezéseket mutatnak. Ezenkívül meghatároztam az egyes topológiai típusok gyakoriságát a ma ismert teljesen annotált baktérium genomokban.

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. A *luxR*, *luxI*, *rsaL* és *rsaM* géneket rendre R, I, L és M betűkkel rövidítem. A génszimbólumok feletti nyíl pedig a transzkripció irányát jelöli.

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 el. 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 három génből álló topológiák viszont sokkal kevésbé változatosak: mindkét esetben egy jellemző topológiát találtam: $\vec{R}\overleftarrow{L}\vec{I}$ és $\overleftarrow{R}\vec{M}\vec{I}$.

A teljesen feltárt baktérium-genomok száma ma már magas, mégis eltörpül az elképzelhető összes baktériumfajhoz képest. Pár általános statisztikai észrevétel azért mégis megtehető: 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.

ID	Minta	Megjelenés a proteobaktériumokban				
		Összes	α	β	γ	δ
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

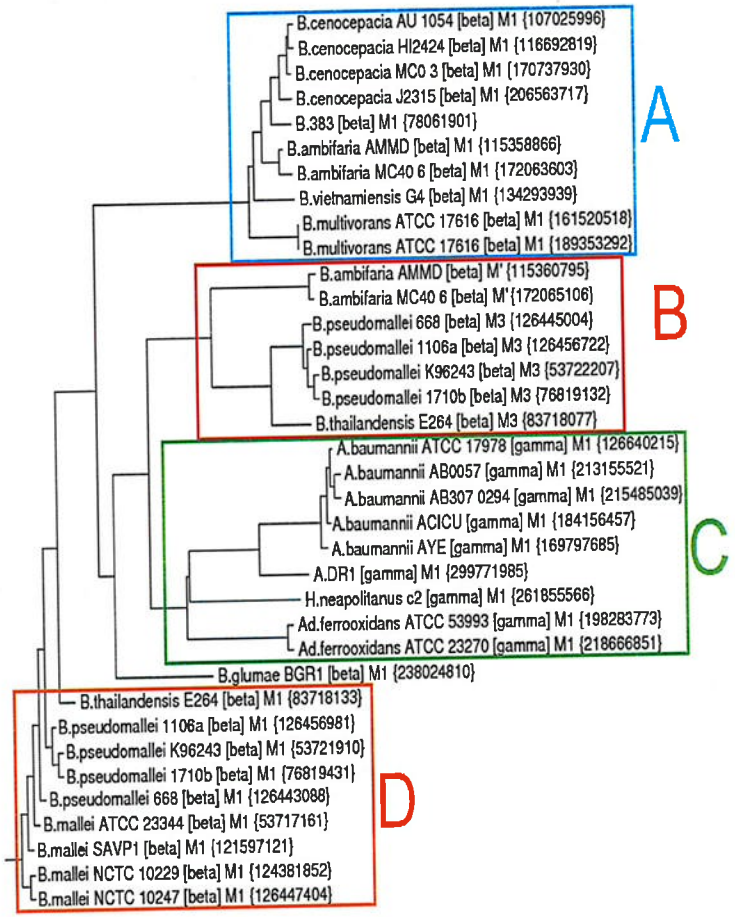
A topológiák gyakoriságát mutató táblázat részlete.

V. Tézis: Megfigyeltem, hogy a quorum sensing szekvenciák a topológiák (és az ezekkel korreláló kémiai jeltípusok) szerint klaszterednek, nem a genomok rokonsága szerint. Vagyis egy bizonyos topológia *luxI* génje jobban hasonlít egy más genom azonos topológiában található génjére, mint a saját genomjában található, de más topológiai típusba tartozó *luxI* génre.

A *LuxI* és *LuxR* fehérjeszekvenciáiból készült cladogrammok elemzése azt mutatta ki, hogy a különböző topológia típusokban szereplő fehérjék tisztán megkülönböztethető csoportokra szeparálódnak. A szabályozófehérjék esetében is ez tapasztalható. Mivel ezek a fehérjék a *quorum sensing*gel rendelkező baktériumok csak egy kisebb csoportjában megfigyelhetőek, ezért könnyebben átlátható és elemezhető ábrákat kaphatunk.

Az *rsaM* génszekvenciák kladogramját megvizsgálva azt tapasztalhatjuk, hogy mind a topológia típusa - amiben a szabályzógén szerepel - mind a baktérium fajok rokonsága számított a fa által kapott csoportok kialakulásában. Az alábbi 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 csoportokban jelentek meg. Röviden tehát a topologiák szerinti csoportok génjei ortológként viselkednek, a köztük lévő viszony pedig paralógiára emlékeztet.



Az rsaM gének 4 csoportja kladogramon ábrázolva

4. Az eredmények alkalmazási területei

A program tervezésekor elsődleges szempont volt a minél általánosabb megvalósítás, ezért az elkészült szoftver nem csak *quorum sensing* rendszerek vizsgálatára alkalmas, hanem lényegében bármely kisméretű alrendszer esetén alkalmazható abban az esetben, ha az azt alkotó gének fehérjetermékei elég nagyszámban ismertek ahhoz, hogy belőlük megfelelő minőségű HMM felismerőket lehessen építeni.

Az általános megfogalmazáson kívül cél volt a program minél teljesebb automatizálása is, így a kezdéshez szükséges adatok összeállítása után a keresés önállóan végrehajtódik, és a futás végén az eredmény elemzésre alkalmas formában jelenik meg. Több keresés is futtatható párhuzamosan, ami ellensúlyozza az algoritmus viszonylag hosszú futási idejét.

Mivel a genomokban nagyon sok a még ismeretlen funkciójú gén, az ilyen jellegű vizsgálatoknak nagyon tág a tere. Dolgozatom befejezése óta csoportunk további 10 kommunkációs rendszer analizisét fejezte be, ezekből egy került publikálásra, a szük keresztmetszetet az emberi feldolgozás jelenti. A kommunikációs gének topológia szerinti csoportosíthatóságát eddig több gén/fehérje családnál sikerült megerősíteni, igy ezen családok ortológiai felosztása a jövőben finomítható lesz.

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 Galbáts Borisz és Erdei Áron segítségét is, 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 bioinmformatika szakterületein.

A szerző publikációi:

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 Gonçalves-Coutinho, 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.

Sanjarbek Hudaiberdiev, K. Sonal Choudhary, Roberto Vera, **Zsolt Gelencsér**, Doriano Lamba and Sándor Pongor."Census of solo luxR genes in becteria" 2014 (előkészületben)