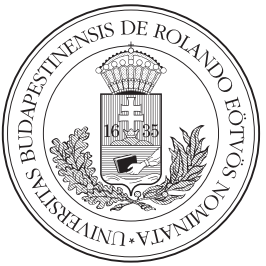
tRNS-ek identitásvizsgálata új, in silico módszerrel

Doktori (PhD) értekezés tézisei

Szenes Áron



Témavezetők:
Dr. Pál Gábor docens és Dr. Jakó Éena tudományos főmunkatárs

Eötvös Loránd Tudományegyetem
Biológia Doktori Iskola
Vezetője: Prof. Dr. Erdei Anna, az MTA levelező tagja

Szerkezeti Biokémia Doktori Program
Vezetője: Prof. Dr. Gráf László, az MTA rendes tagja

2012.

Bevezetés

A teljes élővilágban a fehérjeszintézis egyik kulcsmozzanata a tRNS – aminoacil-tRNS szintetáz kapcsolata. Helyes molekuláris felismerés szükséges ahhoz, hogy a megfelelő tRNS-hez a megfelelő aminosav kapcsolódjon, ezáltal a genomi információ hiba nélkül fejeződjön ki a fehérjében. A tRNS-szintetáz kapcsolatát sok tekintetben feltárták már, azonban sok kérdés megválaszolatlan maradt. Ebben a kapcsolatban szerepet játszó, kiemelt pozíciók az identitáselemek, amelyeket számtalan kísérleti és számítógépes módszerrel igyekeztek feltárni az évek során. A számítógépes módszerekhez először genomi szekvenciákra volt szükség, amelyekből tRNS-detektáló algoritmusok segítségével kellett meghatározni a tDNS szekvenciákat, amelyekből az elmúlt években számtalan adatbázis készült. Számítógépes, funkcionális annotációkra és az identitáselemek in silico meghatározásában már a korábban értek el eredményeket. A másodlagos szerkezetet feltáró algoritmusok és az ezeket használó, közelmúltban fejlesztett adatbázisokban szereplő, másodlagos szerkezeti elemek alapján illesztett tRNS-szekvenciák lehetővé tették újabb predikciós módszerek alkalmazását az identitás meghatározásában. Az információsteória felhasználása is új lehetőségeket nyitott meg nem csak a tRNS szekvenciák vizualizációjában, hanem a determinánsok és antideterminánsok predikciójában is

Az aaRS-okat két osztályba sorolhatjuk szekvenciájuk és térszerkezetük különbözősége alapján. Ezen különbségeknek feltehetően evolúciós okai vannak, amelyek az élővilág mindhárom nagy csoportjában megfigyelhetőek. A szintetázok felosztását követve, annak analógiájára, az általuk aminosavval feltöltött tRNS-eket is besorolhatjuk osztályokba. A tRNS-szintetáz kapcsolat a legtöbb esetben erősen specifikus, ezért feltételezhetjük, hogy egyfajta koevolúciós folyamat során a tRNS szekvenciákban is megmaradt a nyoma annak, hogy melyik osztályba tartozik a szintetáza. Az irodalomban eddig ismert adatok szerint azonban ilyen osztályspecifikus szekvenciaelemek nem léteztek.

Célkitűzések

- Egy új, Jakó Éena által fejlesztett új, diszkrét matematikai megközelítésen alapuló osztályozó módszerrel megdönthetjük-e azt a régi „dogmát”, miszerint a tRNS-ek szekvenciális alapon nem választhatóak ketté a nekik megfelelő, adott aminoacil-tRNS szintetáz osztályuk szerint?
- Ha igen, akkor erre más, Shannon-entrópián alapuló módszer (a „function sequence logo”) is képes-e?
- A végső cél az, hogy tovább növeljük saját eljárásaink hatékonyságát úgy, hogy esetleg új, eddig nem ismert identitáselemekre is javaslatokat lehessen tenni.

Módszerek

Programok, programnyelvek és adatbázisok

A szekvenciák feldolgozásához Unix/Linux „*shell srcipt*”-ek készültek, valamit több esetben az Emboss programcsomag segítségével történtek.
A bonyolultabb feladatokhoz Perl, Python illetve Java programok készültek, a statisztikai analízisek elkészítése és ábrázolása P programnyelven történt.
A tRNS és tDNS szekvenciák az alábbi adatbázisokból származnak: tRNAdb, tRNAdb-CE, SPLITSdb illetve Christian Marck tRNomics adatbázisából.
A szekvencia „*logo*”-k a Makelogo programmal készültek, amely megtalálható a Delila programcsomagban, a „*function logo*”-k a tRNALogofun programmal készültek.

Az „Extended Consensus Partition” (ECP)

A Jakó Éena által fejlesztett ECP módszer működését az alábbi lépésekkel írhatjuk le:

i) Két, többszörösen illesztett szekvenciahalmazból indulunk ki. Az első halmaz a „tanuló” halmaz, amely tartalmazza mind az I. mind a II. osztályból származó, illesztett szekvenciákat, egymástól elkülönítve. A második halmaz egy „kevert” halmaz, amely mindkét osztályból tartalmaz szekvenciákat, de nincsen megjelölve, hogy melyikből.

ii) Az ECP számítása az I. és II. osztályból származó „tanuló” halmaz alapján:

   a) Össze kell gyűjteni azokat a pozíciókat, ahol az egyes osztályhoz zartozó szekvenciák ugyanazt a nukleotidot tartalmazzák. Ezek a nukleotidok az I. illetve a II. osztály „*strictly present*”; elemei.

   b) Meg kell adni minden pozícióra azokat a nukleotidokat, amelyeket az egyes osztályhoz zartozó szekvenciák egyike sem tartalmazza. Ezek a nukleotidok az I. illetve a II. osztály „*strictly absent*”; elemei.

iii) Az osztályozás

   a) A „kevert” adathalmaz minden egyes szekvenciájában egy szekvencia akkor és csak akkor tartozik az adott osztályba, ha:

      1. minden egyes „*strictly present*” elem megtalálható benne; és
      2. minden egyes „ *strictly absent*” elem hiányzik belőle.

Módszerfejlesztés: az AEV

Az ECP azon egyik osztályhoz tartozó *„strictly absent”* elemeit, amelyeket a másik osztály az adott pozícióban tartalmazza, „diszkrimináló elemeknek” (DE) nevezzük.

A továbbfejlesztett ECP analízis során a minden egyes aminosavspecifitású tDNS szekvenciacsoportot minden egyes, tőle különböző aminosavspecifitású tDNS szekvenciacsoporttal összehasonlítva meghatározhatók azok a diszkrimináló elemeket, amelyek arra az aminosavspecifitás csoportra jellemzőek. Az átlagos kizárási érték (*„average excluding value”* – AEV) a DE-k átlaga egy pozícióban.

A fentiek matematikai formalizálásához bevezetjük az $\mathcal{Y}$ változót. Az $\mathcal{Y}$ elemei nukleotidbázisok, tehát $\mathcal{Y} \in \chi$ ahol $\chi = \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$. A változó egyes $\mathcal{Y}_{ik}^{j}$ állapota nem más, mint az a bázis, amelyet egy adott i aminosav-identitás ($i = 1 \ldots, N, N = 20$) j-edik pozíciójában ($j = 1, \ldots, L, L = 96$ – 0-tól a 73-as pozícióig) az identitáshoz tartozó k-adik szekvencia ($k = 1 \ldots, M_i$) tartalmaz. M_i fajonként és aminosav-identitásonként változik.

Bevezethetjük tehát azon bázisok halmazát, amelyek egy i identitás j-edik pozíciójában találhatóak: datok és a

$$\mathcal{Y}_i^j := \{\mathcal{Y}_{ik}^j | k = 1 \ldots, M_i\} \tag{1}$$

Az i aminosav-identitás diszkrimináló elemei (DE) az l aminosavval szemben (szintén: $l = 1 \ldots, N, N = 20$) a j pozícióban:

$$A_{il}^j := \left(\chi \setminus \mathcal{Y}_i^j\right) \cap \mathcal{Y}_l^j \tag{2}$$

Az *„átlagos kizárási érték”* (AEV) számításához a vizsgált j pozícióban defniáljuk az alábbi függvényt:

$$\mathbb{R}\left(A_{il}^j\right) := \begin{cases} 1, & \text{ha } A_{il}^j \neq \varnothing \\ 0, & \text{ha } A_{il}^j = \varnothing \end{cases} \tag{3}$$

Végül a függvény kapott értékeit minden identitás esetén minden identitással szemben összegezzük (tehát összeadjuk azokat az eseteket, amikor találtunk diszkrimináló elemeket a pozícióban), illetve elosztjuk az aminosavak számával:

$$n^j = \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{i=1 \\ i \neq l}}^{N} \mathbb{R}\left(A_{il}^j\right) \tag{4}$$

amely érték nem más, mint az AEV.

Eredmények

1. Statisztikai módszerekkel igazolást nyert, hogy az osztályok szekvencia-alapú szétválasztására az ECP hatékonyabb az eddigi megközelítéseknél.

2. 50 faj (13 ősbaktérium, 30 baktérium és 7 eukarióta) I. és II. osztályú tDNS szekvenciáján elvégezve az ECP analízist, az eddigi módszerekkel nem feltárható, osztályspecifikus DE-ket sikerült meghatározni:

   a) Két, az egész élővilágban általános olyan DE van, amely az I. osztályra jellemző, és amelynek segítségével elkülöníti magát a II. osztálytól: az antikodon középső pozíciója, ahol G35 (az összes fajban) illetve a diszkriminátor bázis, ahol C73 (47 fajban) a DE.

   b) Az élővilágra általános DE-t nem tartalmaz a II. osztály. Azonban jellemző az 1–72 párra a C–G vagy az A–T illetve mind a két féle pár (M–K), mint DE.

   c) A G hiánya a 35. pozícióban erősen jellemző, karakterisztikus elem az I osztályban.

3. Az I. és a II. osztály „inverse function logo"-i reprodukálják az ECP analízis eredményeit.

4. Az egyes pozíciók AEV értékei korrelálnak az eddig publikált identitáselemek számával: az ECP algoritmus képes arra, hogy elkülönítse az identitásoknak megfelelő funkcionális csoportokat.

5. Az algoritmussal feltárhatók olyan pozíciók, amelyek magas AEV értékkel rendelkeznek, emellett még nem írtak le rajtuk ismert identitáselemet. Ezek a pozíciók feltehetően eddig nem ismert identitáselemeket tartalmaznak:

   a) *Escherichia coli* $tRNS^{Trp}$ T31:A39

   b) élesztő $tRNS^{Met}$ T31:T39

   c) élesztő $tRNS^{Ile}$ T30:G40

   d) élesztő $tRNS^{Asp}$ G30:T40 (Ebben az esetben szerkezeti bizpnyíték is van.)

Megjegyzések az eredményekhez

1. Az analízis tDNS szekvenciákon történt, ezért nem tudja figyelembe venni a tRNS szekvenciák poszt-transzkripcionális módosításait.

2. A diszkrimináló elemek (DE) nem csak az identiás kialakításában lehetnek meghatározóak, hanem egyéb más tRNS funkciókban is. Egy nukleotid osztály- vagy identitásspecifikus hiánya eredményezhet olyan tulajdonságot, amely a szerkezeti stabilitáshoz, poszttranszkripcionális módosításhoz, riboszóma vagy elongációs faktor kötéshez szükséges egy tRNS molekulán (néhány példát említve a tRNS összetett funkciói közül).

3. Az analízis szűrt adatbázisokon készült. A szűrés első szempontja az volt, hogy csak megfelelően funkcionáló tRNS-ek legyenek benne: ennek szempontjait a már ismert, minden tRNS-re jellemző nukleotidok jelentették. A másik szempontot a már publikált identitáselemek. Statisztikai módszerekkel bizonyítást nyert azonban, hogy az adatok ilyen szűrése az AEV alapvető karakterisztikáját nem befolyásolják: az eljárás a szűrés elhagyásával is azt eredményezi, hogy a magas AEV értékek zömmel ismert, tehát valódi identitáselemeket hordozó pozíciókon jelennek meg (mint például a diszkriminátor bázis illetve az antikodon tripletje).

4. Az eredményeket azzal a fenntartással kell kezelni, hogy azok akkor nyernek majd igazán értelmet, ha a kísérleti munkák alátámasztják relevanciájukat.

A közlemény alapjául szolgáló publikációk

Áron Szenes, Gábor Pál: *Mapping Hidden Potential Identity Elements by Computing the Average Discriminating Power of Individual tRNA Positions.* DNA Res. 2012 Mar 5. doi: 10.1093/dnares/dss008

Éena Jakó, Péter Ittzés, **Áron Szenes**, Ádám Kun, Eörs Szathmáry, Gábor Pál: *In silico detection of tRNA sequence features characteristic to aminoacyl-tRNA synthetase class membership.* Nucleic Acids Res. 2007;35(16):5593-609.