

Doktori (PhD) értekezés tézisei

**Spliceoszómális iker-intronok (stwintronok)
vizsgálata *Hypoxylon* fonalas gomba
genomokban**

Szabó-Pénzes Fruzsina

Témavezető: prof. Dr. Fekete Erzsébet



DEBRECENI EGYETEM

Juhász-Nagy Pál Doktori Iskola

Debrecen, 2023

1. Bevezetés

A molekuláris biológia és genetika tudományterületeinek fejlődése a XX. század második felétől rohamosan nőtt. Watson és Crick az 1950-es évek elején munkásságukkal mérföldkövet értek el, amikor a DNS kettős szálú spirál felépítését publikálták (1). Néhány nappal e felfedezést követően rájöttek, hogy bázispárosodási szabályoknak megfelelően egy egyszálú DNS alkalmas lehet arra, hogy replikálható legyen, tehát lehetséges a DNS-ről történő másolat készítése (2). A gének mozaikosságát és az intronok jelenlétét Chow és Berget egymástól függetlenül végzett hasonló témájú kutatásaikkal adták tudtunkra 1977-ben, amikor adenovírus kísérleteikkel kimutatták ennek a szekvencia darabnak a létezését (3) (4). Később Gilbert elnevezte ezeket a megszakító szekvenciákat. Az intron nevet kapták az „intra-genic region” kifejezésből eredően, míg a később kifejezésre kerülő mRNS darabokat exonoknak nevezte el (5).

1987-ben Walter Gilbert megírta a „The Exon Theory in Genes” című cikkét (6), amiben a „korai intron” teóriát állítja fel, miszerint az intronok már az első gének létrejöttkor léteztek. Ezzel ellentétben a másik oldal a „késői intron” teóriát látja igazabbnak. Eszerint az intronok az evolúció

során, később kerültek be az exonok közé (7) (8). Az intronok különbözőségük miatt csoportokba oszthatók. Ez alapján a korábban felállított mindkét intron teória igaznak bizonyulhat. A prokarióták génjeiben ribozim intronok lehettek jelen, mutatja ezt az evolúciós megfigyelés, és a mitokondriális, valamint kloroplasztok DNS-ében jelenlévő intron tartalom. Azonban később spliceoszómális intronok alakultak ki (9).

Copertino és Hallick voltak, akik először írtak egy speciális intron struktúráról, a twintronokról 1991-ben. Ez nem csak a génkifejeződésben jelent egy újabb evolúciós lépést, de jelenlegi kutatásom alappilléret is adta. Az általuk leírt szekvencia két intron egymásba ágyazódásaként volt jelen az exonok között. A twintron szót az angol „twin intron” szavakból alkották meg (10).

A doktori disszertációm témaköre a twintronokhoz fűződik. A Debreceni Egyetem Biomérnöki Tanszékének twintronokról szóló első publikációja 2013-ban volt (11). E kutatás során a twintronok között is egy speciálisabb szekvenciát véltek felfedezni, amit stwintronoknak neveztek el. A stwintronok olyan „intron-az-intronban” megszakító szakaszok, amiket csak spliceoszómális intronok alkotnak, ebből adódóan az elnevezésük a spliceosomal twintron szavakból ered. Kezdetben a tanszéki kutatások célja minél

több stwintron igazolása volt különböző fonalas gomba fajokban, amit több tanszéki cikk is bizonyít (12) (13) (14) (15). Ehhez hozzájárult a jelenlegi értekezés alapjául szolgáló valamennyi cikk is, melyeket a tanszék kutatócsoportjának tagjaként mutathatok be (16) (17) (18). Egyre több fajban és egyre érdekesebb funkcióval rendelkező stwintronokat találunk. Segítségükkel megfigyelhetjük a gombák evolúciós változását és felfedhetjük filogenetikai tulajdonságaikat is.

2. Célkitűzések

A kutatócsoportunk célja az első stwintron struktúra feltárását követően kezdetben minél több stwintron leírása volt. Az évek során számos fonalas gomba genomot vizsgáltunk át és sikeresen azonosítottunk stwintronokat. Így ez alkalommal a következő célokat tűztük ki a tervezett kísérletsorozathoz:

- I. Fonalas gomba genomtól eltérő, élesztőgomba fajokban található stwintronok igazolása molekuláris biológiai módszerekkel.
- II. A stwintron kereső modellünk átdolgozása és finomhangolása [D1,2] típusú struktúrákra.

- III. Több ugyanazon [D1,2] típusú stwintronokat tartalmazó genom keresése, majd igazolása laboratóriumi körülmények között és *in silico* analízissel.
- IV. Az összegyűjtött stwintronok szekvencia elemzése, szimmetriai vizsgálata, valamint a másodlagos szerkezetének lehetséges felépítésének meghatározása.
- V. Az stwintronok lehetséges stwintronizációs folyamatainak kikövetkeztetése, feltárása.
- VI. Tényezők meghatározása, amelyek befolyásolhatják és/vagy kialakíthatják a stwintronok proliferációját.

3. Anyagok és módszerek

3.1 Bioinformatikai módszerek

1.1.1 Intron modellek, potenciális stwintronok keresése

A stwintronok keresésére a Kupfer által publikált spliceoszómális intronokra jellemző modellt (19) alkalmaztuk. Kupfer *A. nidulans*-ra alkotott modelljét 2017-es cikkünkben kezdtük el összetett intronokra alkalmazni, amikor a lariat első betűjelét, „R”-t „D”-re és az akceptor szekvencia első betűjelét „Y”-ról „H”-ra cseréltük, ezzel bővítve a potenciális intron jelöltek számát a keresés során. Ezt a modellt a külső és belső intronokra egyaránt használtuk (13). A spliceoszómális intronok konzervatív részei: a donor

első két nukleotidja: GT; a lariat elágazási pontja: A; és az akceptor utolsó két nukleotidja: AG. A konszenzus szekvenciák közötti távolságok meghatározásához számításba vettük az alábbi szabályokat: a genom átlagos intron hossza 73 nt, a minimum intronhossz 43 nt, a maximum intronhossz 160 nt. A donor és lariat közötti szekvencia szakasz nagyobb, mint a lariat és akceptor közötti, így a donor és lariat között 25-120 bázispár, a lariat és akceptor között 4-24 bázispár távolságot állítottunk be (13) (19).

A modelleket [D1,2] és [D5,6] stwintronokra alkottuk meg, valamint további [D1,2] stwintronok keresésére további módosításokat végeztünk. A külső- és belső intronokhoz tartozó donorok 3. 4. és 6. nukleotidját „N”-re cseréltük, ami mind a négy nukleotid lehetőséget engedi 5'-GTNNGN. Azonban az összes, 64 lehetséges intron variációból csak 24 félélet alkalmaztunk, aminek az alapja egy korábbi, nem publikált eredmény. Az eredmények könnyebb kezelhetősége érdekében kizártuk a kevésbé kanonikus 5'-GC típusú donorokat és az 5'-GAG akceptor szekvenciákat a találatok közül.

3.1.1 Génbányászat

A sister (nagy hasonlóságot mutató) [D1,2] stwintronok gyűjtése a *Hypoxylon sp. CO27-5* genomjából történt (Gene Bank Master Accession number MDCL00000000) BLASTN használatával a NCBI (National Center for Biotechnology Information) hivatalos online adatbázisában. Az *A. nidulans* (GenBank AACD01000130, CDS koordináták: start-stop: 208790-210069, publikálatlan munka) egy korábbi kutatásból származó jól konzervált mitokondriális szállító fehérje ortológjainak keresésére TBLASTN segítségével Genome Shotgun contig (WGS) adatbázisokat kutattunk át. A *Xylariales* rendbe és *Hypoxylaceae* családba tartozó fajok közül ortológ és paralóg gének egyaránt kigyűjtésre kerültek.

Az intron-exon struktúrákat manuálisan határoztuk meg és minden génre végeztünk analízist, így az összes szekvenciaeredetű információt összegyűjtöttük. A *Hypoxylon sp. CO27-5*, *Hypoxylon sp. EC38* és *Daldinia sp. EC12* legtöbb génmodelljét SRA adatokkal megerősítettük, ugyanis az NCBI Sequence Read Archive (SRA) a legtöbb stwintron köztes szekvenciáját (splicing intermedier, „splinter”) igazolja. Funkcionálisan annotáltuk a legtöbb

sister stwintront hordozó gént a Protein Families (Pfam) adatbázisból (20).

3.1.2 Filogenetikai analízis és egyéb informatikai módszerek

A Multiple sequence Alignment with Fast Fourier Transform (MAFFT: version 7) (21) programmal többszörös illesztéseket végeztünk. Az informatív találatok a Block Mapping és Gathering with Entropy (BMGE, version 1.12) (22) szoftver segítségével került kiválasztásra. A maximum likelihood törzsfákat PhyML (version 3.0) program automatikus szubsztitúciós modelljét alapértelmezett módban (23) vagy a Smart Model Selection (SMS) szoftvert alkalmaztuk szubsztitúciós modell választásra (24) az Akaike Information Criterion (AIC) becslését használva. Egy másik esetben pedig az SMS-PhyML 5000 ismétlést futtatott Felsenstein-féle bootstrap (25) alapján az ágak kikövetkeztetéséért. A fa ágainak hosszát az Approximate likelihood-ratio test (26) vagy a Shimodaira-Hasegawa-like approximate likelihood-ratio tests (SH-like aLRT) (26) biztosította. A törzsfa rajzolása FigTree (version 1.4.3) program segítségével történt a PhyML Newick kimeneteléből. A további szerkesztéseket Adobe Illustrator (Adobe Inc., San Jose, CS, USA) programmal készítettük.

Két vagy három intronikus szekvencia összehasonlítására a Clustal Omega (27) szekvenciaillesztését alkalmaztuk. A fordított terminális ismétlődésekkel rendelkező egyszálú elágazó RNS intronok másodlagos szerkezetének a kalkulálásához az RNAfold (28), míg a teoretikus hibridizációs energia- és bázispárosodásból következő dupla-szálú molekula kalkulációja ugyan azokból a (stw)intron RNS-ekből az RNAcofold (28) (29) programmal történt. A sister (stw)intronok pre-mRNS-einek 5'- és 3'-exon-intron találkozásainak közvetlen környezetének szekvencia logóját a Skyline (30) program adta alapértelmezett beállításokkal. Az RNAfold (29) (28) segítségével képeztük a másodlagos szerkezeteket az izolált bázispárok elkerülése nélküli beállítás mellett, és a legkisebb minimum szabad energia felhasználási adatokat is feljegyeztük (ΔG). Ezeket összehasonlítottuk az Mfold (UNAFold weboldal) (31) (32) által kapott alternatív eredményekkel. A sister stwintronokkal összehasonlítást végezve, a másodlagos szerkezet javítására a MAFFT eredményeken manuálisan korrigálni kellett egy-két sister stwintron esetében. Ezek másodlagos szerkezetét RNAalifold (ViennaRNA webhely) (33) adta.

3.2 Gombatörzsek

Lipomyces suomiensis CBS 7251 (Westerdijk Fungal Biodiversity Institute, Royal Netherlands Academy of Sciences, Utrecht, The Netherlands) (34), *Lipomyces kononenkoae* CBS 2514 (34), *Lipomyces lipofer* CBS 944 (National Collection of Agricultural és Industrial Microorganisms, Budapest, Hungary: NCAIM Y.00351) (34), *Lipomyces starkeyi* CCY 33-1-1 (Dept. of Microbiology és Biotechnology, Corvinus University of Budapest, Hungary) (35) és *Hypoxyton* sp. CO27-5 (*Ascomycota*; *Pezizomycotina*; *Sordariomycetes*; *Xylariales*; *Hypoxyloaceae*; *Hypoxyton*, besorolatlan) (36).

4. Új tudományos eredmények

- I. 3 új típusú stwintron [D4,5]; [D6,7]; [D7,8] igazolása *Lipomyces* élesztő gomba fajokban.
- II. Új [D5,6]- és [D1,2] kereső modell megalkotása, valamint 4 további kibővített [D1,2] modell módosítás.
- III. 117 [D1,2] típusú stwintron felfedése *Hypoxyloaceae* családon belüli fajok esetében.
- IV. Azonos típusú [D1,2] stwintronok elemzése közeli fonalas gomba genomokban.

- V. Új lehetséges stwintronizációs és intronizációs folyamatok feltárása.
- VI. Felállítottunk egy hipotézist, miszerint a magas szintű szimmetriai jellem és TIR régiók elősegíthetik a stwintronok terjedését.

5. Összefoglalás

A stwintronok (spliceoszómális iker-intronok) két spliceoszómális intronból épülnek fel úgy, hogy az egyik intron a másik intron egy konszenzus elemébe (donor, lariat vagy akceptor) van beépülve. A struktúra kivágódása kétlépéses folyamat, ugyanis a külső intron nem tud kivágódni, amíg a belső intron megszakítja.

Lipomyces élesztő fajok retukulum fehérjét kódoló génjeiben 3 új típusú stwintront, fedeztünk fel: [D4,5], [D6,7] és [D7,8]. Emellett egy stwintronizációs folyamatot is felfedtünk. Az eredeti intron donorja (vagy egy része) megkettőződhetett, ezzel egy új vágási helyet kialakítva. A DSB mechanizmusnak köszönhetően, amit a cNHEJ folyamat kísér, létre jön egy új 5' vágási hely, ami a belső intron donorját fogja alkotni. A belső intron 3' végi konszenzus helyei pedig pontmutációk révén alakulhattak ki.

[D1,2] típusú stwintronokra bukkantunk *Hypoxylon sp.* CO27-5 és közeli rokonjainak ortológ génjeiben. A keresőmodellünket addig alakítottuk és szerkesztettük, amíg végül több, mint 100 [D1,2] típusú stwintront találtunk a CO27-5 fajon belül és közeli rokonjaiban egyaránt. Az összesen 117 stwintront, két csoportra osztottuk: „sister” (többszörösen előforduló; MO) és „nem-sister” (egyedileg előforduló; UO). A sister stwintronok magas szintű homológiát mutattak egymással, összesen 36 [D1,2] szekvencia tartozik ide (23 teljes sister, 13 perciálisan sister). Emellett találtunk 12 kanonikus intront is (1-es- és 2-es típusú) melyek hasonlóságot mutatnak a sisterekhez. 10 esetben a stwintron 5' végi belső intron szekvenciájának és a 3' végi külső intron szekvenciájának a fúziójaként véve mutat magas szintű hasonlóságot. Ez alapján a sister stwintronok/intronok csoportja egy újfajta intronizációs folyamatot követhet, ugyanis a stwintronokban található két palindrom szekvenciájának rekombinációja a stwintron középső részének kivágódását okozza, ezzel létrehozva egy középső részen palindromot (5'-TTTCTAGAAA) tartalmazó kanonikus intront (2-es típus). Egy 2-es típusú cropped intron (HCOc121A) esetében az intron inszerciója 3 nukleotid deléciójához köthető, ami alapján bemutattunk egy lehetséges proliferációt is.

Vizsgáltuk a 81 UO [D1,2] stwintront is, valamint összehasonlítottuk őket a sister stwintronokkal. A stwintront tartalmazó géneket elemeztük és felkutattuk a közeli és távolabbi rokon fajok ortológ és paralóg génjeiket is. Ezzel több, mint 1500 stwintron volt érintett az összehasonlító elemzésben. A sister stwintronok fiatalabb szekvenciák, ezt bizonyítja, hogy a *CO27-5* mellett, csak nagyon közeli rokonjában, az *EC38*-ban fordulnak elő. Ezzel szemben a 81 UO stwintron távolabbi rokon fajokban is megjelenik. A két csoportjában leginkább az egyes fázis dominál. A sister stwintronok külső intronjának BP és akceptor közötti része háromszor olyan hosszú (19 nt), mint az idősebb UO stwintronoknál (5-7 nt). Mind a 117 [D1,2] stwintron esetében elmondható, hogy az AU tartalom aránya magasabb. A sister stwintronok csoportjának az % AU értéke viszont lényegesen magasabb, mint a 81 UO stwintronból álló csoportnak. Belső szimmetria ugyanúgy jellemzi mindkét csoportot, azonban a sister stwintronok a láncvégi TIR régiók miatt stabilabb másodlagos szerkezetet eredményeznek. A sister stwintronok ΔG értéke átlagosan 18–20 %-kal alacsonyabb, mint az azonos mérettartományban lévő UO stwintronok értékei.

6. Irodalomjegyzék

1. Molecular structure of nucleic acids. Watson, James D. és Crick, Francis H. C. 1953. 04 25, Nature, 171.:737-738.
2. Genetical Implications of the Structure of Deoxyribonucleic Acid. Watson, James D. és Crick, Francis H. C. 1953. 05 30, Nature, 171.:964-967.
3. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Chow, Louise T., és mtsai. 1977. 09 01, Cell, 12.:1-8.
4. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Berget, S. M., Moore, C. és Sharp, P. A. 8, 1977. 08, Proc Natl Acad Sci U S A., 74.:3171-3175.
5. Why genes in pieces? Gilbert, Walter. 1978. 02 01, Nature, 271.:501.
6. The Exon Theory of Genes. Gilbert, Walter. 1987., Cold Spring Harb Symp Quant Biol., 52. : 901-905.
7. The recent origins of introns. Palmer, Jeffrey D. és Logsdon, John M. Jr. 4, 1991. 12, Current Opinion in Genetics & Development, 1.:470-477.
8. Origin of Genes. Gilbert, Walter, Souza, Sésro J. de és Long, Manyuan. 15, 1997. 07 22, Proceedings of the National Academy of Sciences, 94.:7698-7703.
9. The origin of introns és their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Koonin, Eugene V. 22, 2006. 08 14, Biology Direct, 1. .
10. Group II twintron: an intron within an intron in a chloroplast cytochrome b-559 gene. Copertino, Donald W. és Hallick, Richard B. 2, 1991., The EMBO Journal, 10.:433-442.

11. Spliceosome twin introns in fungal nuclear transcripts. Flippi, Michel, és mtsai. 2013. 01 19, *Fungal Genetics és Biology*, 57.:48-57.
12. Alternatively spliced, spliceosomal twin introns in *Helminthosporium solani*. Ág, Norbert, és mtsai. 2015. 10 26, *Fungal Genetics és Biology*, 85.:7-13.
13. A mechanism for a single nucleotide intron shift. Fekete, Erzsébet, és mtsai. 15, 2017. 06 08, *Nucleic Acids Research*, 45.:9085-9092.
14. Emergence és loss of spliceosomal twin introns. Flippi, Michel, és mtsai. 7, 2017. 10 06, *Fungal Biology és Biotechnology*, 4.
15. A spliceosomal twin intron (stwintron) participates in both exon skipping és evolutionary exon loss. Kavalecz, Napsugár, és mtsai. 9940, 2019. 07 09, *Scientific Reports*, 9.
16. Complex intron generation in the yeast genus *Lipomyces*. Ág, Norbert, és mtsai. 6022, 2020. 04 07, *Scientific Reports*, 10.
17. Internally Symmetrical Stwintrons és Related Canonical Introns in Hypoxylaceae Species. Fekete, Erzsébet, és mtsai. 710, 2021. 08 29, *Journal of Fungi*, 7.
18. Unique és Repeated Stwintrons (Spliceosomal Twin Introns) in the Hypoxylaceae. Fekete, Erzsébet, és mtsai. 397, 2022. 04 13, *Journal of Fungi*, 8.
19. Introns és Splicing Elements of Five Diverse Fungi. Kupfer, Doris M., és mtsai. 5, 2004. 10 3, *Eukaryotic Cell*, 3. : 1088–1100.
20. The Pfam protein families database in 2019. El-Gebali , Sara, és mtsai. D1, 2019., *Nucleic Acids Res*, 47.:D427-D432.
21. MAFFT multiple sequence alignment software version 7: improvements in performance és usability. Katoh, Kazutaka és Stésley, Daron M. 4, 2013., *Mol Biol Evol*, 30.:772-780.

22. BMGE (Block Mapping és Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. Criscuolo, Alexis és Gribaldo, Simonetta. 210, 2010., *BMC Evol Biol*, 10.
23. SMS: Smart Model Selection in PhyML. Lefort , Vincent, Longueville, Jean-Emmanuel és Gascuel, Olivier. 9, 2017., *Mol Biol Evol*, 34. : 2422-2424.
24. An improved general amino acid replacement matrix. Le, Si Quang és Gascuel, Olivier. 7, 2008., *Mol Biol Evol*, 25. : 1307-1320.
25. Confidence limits on phylogenies: An approach using the bootstrap. Felsenstein, Joseph. 4, 1985., *Evolution*, 39. : 783-791.
26. Approximate likelihood-ratio test for branches: A fast, accurate, és powerful alternative. Anisimova, Maria és Gascuel, Olivier. 4, 2006., *Syst Biol*, 55.:539-552.
27. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Sievers, Fabian, és mtsai. 539, 2011., *Mol Syst Biol*, 7.
28. The Vienna RNA websuite. Gruber , Ésreas R., és mtsai. 2008., *Nucleic Acids Res*, 36. : W70-4.
29. Partition function és base pairing probabilities of RNA heterodimers. Bernhart, Stephan H., és mtsai. 3, 2006., *Algorithms for Mol Biol*, 1.
30. Skylign: a tool for creating informative, interactive logos representing sequence alignments és profile hidden Markov models. Wheeler, Travis J., Clements, Jody és Finn , Robert D. 7, 2014., *BMC Bioinform.*, 15.
31. Mfold web server for nucleic acid folding és hybridization prediction. Zuker, Michael. 13, *Nucleic Acids Research* : ismeretlen szerző, 2003., 31.:3406–3415.

32. UNAFold: software for nucleic acid folding és hybridization. Markham, Nicholas R. és Zuker, Michael. 2008., *Methods Mol Biol*, 453. : 3-31.
33. RNAalifold: improved consensus structure prediction for RNA alignments. Bernhart, Stephan H., és mtsai. 474, 2008., *BMC Bioinformatics*, 9.
34. Tempo és mode of genome evolution in the budding yeast subphylum. Shen, Xing-Xing, és mtsai. hely nélk. : 175, 2018., *Cell*, 6. : 1533–1545.
35. Comparative genomics of biotechnologically important yeasts. Riley, Robert, és mtsai. 35, 2016., *PNAS*, 113. : 9882-9887.
36. Characterization of four endophytic fungi as potential consolidated bioprocessing hosts for conversion of lignocellulose into advanced biofuels. Wu, Weihua, és mtsai. 6, 2017., *Appl Microbiol Biotechnol*, 101. : 2603-2618.



Nyilvántartási szám: DEENK/82/2023.PL
Tárgy: PhD Publikációs Lista

Jelölt: Szabó-Pénzes Fruzsina
Doktori Iskola: Juhász-Nagy Pál Doktori Iskola
MTMT azonosító: 10069262

A PhD értekezés alapjául szolgáló közlemények

Idégen nyelvű tudományos közlemények külföldi folyóiratban (3)

1. Fekete, E., **Pénzes, F.**, Ág, N., Ág-Rácz, V., Karaffa, E. M., Scazzocchio, C., Flipphi, M., Karaffa, L.: Unique and Repeated Stwintrons (Spliceosomal Twin Introns) in the Hypoxylaceae. *J. Fungi*. 8 (4), 1-23, 2022. EISSN: 2309-608X.
DOI: <http://dx.doi.org/10.3390/jof8040397>
IF: 5.724 (2021)
2. Fekete, E., **Pénzes, F.**, Ág, N., Scazzocchio, C., Flipphi, M., Karaffa, L.: Internally Symmetrical Stwintrons and Related Canonical Introns in Hypoxylaceae Species. *J. Fungi*. 7, 1-21, 2021. EISSN: 2309-608X.
DOI: <http://dx.doi.org/10.3390/jof7090710>
IF: 5.724
3. Ág, N., Kavalecz, N., **Pénzes, F.**, Karaffa, L., Scazzocchio, C., Flipphi, M., Fekete, E.: Complex intron generation in the yeast genus *Lipomyces*. *Sci. Rep.* 10, 1-10, 2020. EISSN: 2045-2322.
DOI: <http://dx.doi.org/10.1038/s41598-020-63239-6>
IF: 4.379

A közlő folyóiratok összesített impakt faktora: 15,827

A közlő folyóiratok összesített impakt faktora (az értekezés alapjául szolgáló közleményekre): 15,827

A DEENK a Jelölt által az iDEa Tudóstérbe feltöltött adatok bibliográfiai és tudománymetriai ellenőrzését a tudományos adatbázisok és a Journal Citation Reports Impact Factor lista alapján elvégezte.

Debrecen, 2023.03.16.



**Short thesis for the degree of doctor of philosophy
(PhD)**

**Investigation of spliceosomal twin-introns
(stwintrons) in the genomes of filamentous
fungus *Hypoxylon***

by Fruzsina Szabó-Pénzes

Supervisor: prof. Dr. Erzsébet Fekete



UNIVERSITY OF DEBRECEN

Juhász-Nagy Pál Doctoral School

Debrecen, 2023

1. Introduction

The development of the scientific fields of molecular biology and genetics has increased from the the second half of the 20th century. Watson and Crick achieved a milestone with their work in the early 1950s when they published the structure of the DNA double-stranded helix (1). A few days after this discovery, they realized that according to base-pairing rules, a single-stranded DNA can be capable of being replicated, so it is possible to make a copy of DNA (2). The mosaicism of genes and the presence of introns were revealed to us by Chow and Berget independently in 1977, when they demonstrated the existence of this sequence with their adenovirus experiments (3) (4). Later, Gilbert named these interrupting sequences, derived from the term "intragenic region", while the expressed mRNA pieces were named exons (5).

In 1987, Walter Gilbert wrote his article "The Exon Theory in Genes" (6), in which he proposed the "early intron" theory. This theory says introns already existed when the first genes were appeared. On the other hand, the other side considers the "late intron" theory. According to this theory, introns were later added to exons during evolution (7) (8). Introns can be divided into groups due to their diversity. so

both previously proposed intron theories may prove to be true. Ribozyme introns could be present in the genes of prokaryotes, this is shown by evolutionary observations and the intron content present in mitochondrial and chloroplast DNA. Later, however, spliceosomal introns arose (9).

In 1991, Copertino and Hallick were the first to write about a special intron structure, twintrons. This not only represents another evolutionary step in gene expression, but also provided the foundation of my current research. The sequence they described was present as a nesting of two introns between exons. The word twintron was created from the words "twin intron" (10).

The subject of my doctoral dissertation is related to twintrons. The first publication about twintrons by the Department of Biochemical Engineering of the University of Debrecen was in 2013 (11). During the research, they discovered a special sequence among the twintrons, which they named stwintrons. Stwintrons are "intron-within-intron" intervening sequences formed only by spliceosomal introns, hence their name from the words spliceosomal twintron. In the beginning, the goal of the departmental research was to verify as many stwintrons as possible in various species of filamentous fungi, which is proven by several departmental articles (12) (13) (14) (15). With this thesis I can present as a

member of the department's research group 3 more article about stwintrons (16) (17) (18). We come across stwintrons in more and more species and with more and more interesting functions. With this investigation, we can observe the evolutionary change of fungi and reveal their phylogenetic properties.

2. Objectives

After the discovery of the first stwintron structure, the goal of our research group was initially to describe as many stwintrons as possible. Over the years, we have screened many filamentous fungi genomes and successfully identified stwintrons. Thus, this time we set the following goals for the planned series of experiments:

- I. Confirmation of stwintrons found in yeast species by molecular biological methods.
- II. Revision and fine-tuning of our stwintron search model for [D1,2] type structures.
- III. Search for genomes containing the same [D1,2] type stwintrons and verify them in laboratory conditions and by in silico analysis.

- IV. Analysis of the collected stwintron sequences in structure, sequentially, in symmetry and by examining the secondary structure.
- V. Reveal the possible stwintronization processes of the examined stwintrons.
- VI. Determination of factors affecting the proliferation of stwintrons.

3. Materials and methods

3.1 Bioinformatic methods

3.1.1 Intron models, searching tool for putative stwintrons

The basic model from which we started is the model for spliceosomal introns published by Kupfer. We began to apply Kupfer's model for *A. nidulans* to stwintrons in our 2017 paper. We changed the first letter of the lariat from "R" to "D" and the first letter of the acceptor sequence from "Y" to "H", thus expanding the number of potential stwintron candidates in the search. We used this model for both external and internal introns (13). Conservative parts of spliceosomal introns: first two nucleotides of the donor: GT; branching point of the lariat: A; and the last two nucleotides of the acceptor: AG. The following rules were used to

determine the distances between consensus sequences: the average intron length of the genome is 73 nt, the minimum intron length is 43 nt, and the maximum intron length is 160 nt. The sequence section between donor and lariat is larger than between lariat and acceptor, so we set a distance of 25-120 base pairs between donor and lariat, and 4-24 base pairs between lariat and acceptor (13) (19).

We created models for [D1,2] and [D5,6] stwintrons in these experiments. Furthermore, we made modifications to search for additional [D1,2] stwintrons. The 3rd, 4th and 6th nucleotides of the donors belonging to the external and internal introns were changed to "N", which allows all four nucleotides in these positions (5'-GTNNGN). However, we used only 24 of the 64 possible intron variations, which is based on a previous, unpublished result. In order to make the results easier to handle, we excluded less canonical 5'-GC type donors and 5'-GAG acceptor sequences from the results.

3.1.2 Genome mining

The sister (high similarity) [D1,2] stwintrons were collected from genome of *Hypoxylon sp. CO27-5* (Gene Bank Master Accession number MDCL00000000) using BLASTN in the official online database of NCBI (National Center for Biotechnology Information). We searched for orthologs of a

well-conserved mitochondrial transporter protein from *A. nidulans* (GenBank AACD01000130, CDS coordinates: start-stop: 208790-210069, unpublished work) in Genome Shotgun contig (WGS) databases using TBLASTN. Both orthologous and paralogous genes were collected from species belonging to the order *Xylariales* and the family *Hypoxylaceae*.

The intron-exon structures were defined manually and an analysis was performed for each gene. Most of the gene models of *Hypoxylon sp. CO27-5*, *Hypoxylon sp. EC38* and *Daldinia sp. EC12* were confirmed with SRA data. The NCBI Sequence Read Archive (SRA) verifies the splicing intermediate (splinter) of most stwintrons. We functionally annotated most sister stwintron-carrying genes from the Protein Families (Pfam) database (20).

3.1.3 Phylogenetic analysis and other informatic methods

Multiple alignments were performed with the Multiple sequence Alignment with Fast Fourier Transform (MAFFT: version 7) (21) program. Informative hits were selected using Block Mapping and Gathering with Entropy (BMGE, version 1.12) (22) software. In one case we used the PhyML (version 3.0) program in default mode (23) with the automatic

substitution model. Other case we used the Smart Model Selection (SMS) software for substitution model selection (24). The Akaike Information Criterion (AIC) estimate for the maximum likelihood trees. SMS-PhyML ran 5000 replicates based on Felsenstein's bootstrap (25) to infer the branches. The length of the branches of the tree was provided by the Approximate likelihood-ratio test (26) or Shimodaira-Hasegawa-like approximate likelihood-ratio tests (SH-like aLRT) (26). The stem tree was drawn using the FigTree (version 1.4.3) program from the PhyML Newick output. Further editing was done with Adobe Illustrator (Adobe Inc., San Jose, CS, USA).

The Clustal Omega (27) sequence alignment helped us to compare two or three intronic sequences. RNAfold (28) is used to calculate the secondary structure of single-stranded branched RNA introns with inverted terminal repeats, while the calculation of the double-stranded molecule resulting from the theoretical hybridization energy and base pairing from the same (stw)intron RNAs is RNAcifold (28) (29) program. The sequence logo of the immediate vicinity of the 5'- and 3'-exon-intron junctions of the pre-mRNAs of the sister (stw) introns was provided by the program Skyline (30) with default settings. RNAfold (29) (28) was used to form the secondary structures with the isolated base pairs set without

evasion. The smallest minimum free energy usage data was also recorded (ΔG). These were compared with alternative results obtained by Mfold (UNAFold website) (31) (32). The MAFFT results had to be corrected manually in the case of one or two sister twintrons. Their secondary structure was provided by RNAalifold (ViennaRNA website) (33).

3.2 Fungal strains

Lipomyces suomiensis CBS 7251 (Westerdijk Fungal Biodiversity Institute, Royal Netherlands Academy of Sciences, Utrecht, The Netherlands) (34), *Lipomyces kononenkoae* CBS 2514 (34), *Lipomyces lipofer* CBS 944 (National Collection of Agricultural and Industrial Microorganisms, Budapest, Hungary: NCAIM Y.00351) (34), *Lipomyces starkeyi* CCY 33-1-1 (Dept. of Microbiology and Biotechnology, Corvinus University of Budapest, Hungary) (35) and *Hypoxylon sp. CO27-5* (*Ascomycota; Pezizomycotina; Sordariomycetes; Xylariales; Hypoxylaceae; Hypoxylon*, unclassified) (36).

4. New scientific results

- I. Confirmation of 3 new types of twintrons [D4,5]; [D6,7]; [D7,8] in *Lipomyces* yeast species.

- II. Creation of a new [D5,6]- and [D1,2] search model, and making 4 additional extended [D1,2] model modifications.
- III. Discovery of 117 [D1,2]-type stwintrons in species within the family *Hypoxylaceae*.
- IV. Analysis of same-type [D1,2] stwintrons in closely related filamentous fungal genomes.
- V. Exploration of new possible stwintronization and intronization processes.
- VI. We hypothesized that the high symmetry character and TIR regions can promote the propagation of (stw)introns.

5. Summary

Stwintrons (spliceosomal twin introns) are consist of two spliceosomal introns, An intron inserted into a consensus element (donor, lariat or acceptor) of another intron. Excision of this structure is a two-step process, as the external intron cannot be excised until it is interrupted by the internal intron.

In genes encoding reticulon-like proteins of *Lipomyces* yeast species, we discovered 3 new types of twintrons: [D4,5], [D6,7] and [D7,8]. We were also able to reveal a stwintronization process. The donor (or a part) of the original

intron could be duplicated and created a new splicing site. The DSB mechanism with the cNHEJ process form a new 5' splicing site that will become the donor of the internal intron. The consensus sites at the 3' end of the internal intron could have formed through point mutations.

We found more than 100 [D1,2] type stwintrons both within *the Hypoxylon sp. CO27-5* species and within its close relatives. We divided them into two groups: "sister" (multiple occurring; MO) and "non-sister" (uniquely occurring; UO) stwintrons. The sister stwintrons showed a high level of homology with each other, including a total of 36 [D1,2] sequences (23 full sisters, 13 partial sisters). In addition, we found 12 canonical introns (type-1 and type-2) that show similarity to sisters. In 10 instances, type-2 cropped sister introns consist of a fusion of 5' sequences of the internal intron with 3' sequences of the external intron. Based on this observation, the group of sister stwintrons/introns can also follow a new kind of stwintronization process, because the recombination of the two palindrome sequences in the stwintrons causes the middle part of the stwintron to be excised, thereby creating a canonical intron containing a palindrome (5'-TTTCTAGAAA) in the middle (type-2). In one type-2 cropped intron (HCOc121A) the insertion of the intron can be linked to the deletion of 3 nucleotides and we

use this observation to demonstrate proliferation of (stw)introns.

We also investigated the 81 UO [D1,2] stwintrons and compared them with sister stwintrons. We analyzed all genes containing stwintrons and searched for their orthologous and paralogous genes in closely and more distantly related species. With this study, more than 1,500 stwintrons were tested in the comparative analysis. The sister stwintrons are younger sequences, they occur mostly in its very close relative, EC38. In contrast, the 81 UO stwintron also appears in more distantly related species. Phase 1 dominates the two groups of stwintrons. The part of the external intron of the sister stwintrons between BP and acceptor is three times as long (19 nt) as in the older UO stwintrons (5-7 nt). In the case of all 117 [D1,2] twintrons we can say that the ratio of AU content is higher, and the % AU of the group of sister swintrons is significantly higher than that of the group of 81 UO swintrons. Internal symmetry characterizes both groups in the same way, however, sister twintrons result in a more stable secondary structure due to the TIR regions. The ΔG value of the sister stwintrons is on average 18–20% lower than the UO in the same size range values of stwintrons.

6. References

1. Molecular structure of nucleic acids. Watson, James D. and Crick, Francis H. C. 1953. 04 25, *Nature*, 171.: 737-738.
2. Genetical Implications of the Structure of Deoxyribonucleic Acid. Watson, James D. and Crick, Francis H. C. 1953. 05 30, *Nature*, 171.: 964-967.
3. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Chow, Louise T., et al. 1977. 09 01, *Cell*, 12.:1-8.
4. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Berget, S. M., Moore, C. and Sharp, P. A. 8, 1977. 08, *Proc Natl Acad Sci U S A.*, 74.3171-3175.
5. Why genes in pieces? Gilbert, Walter. 1978. 02 01, *Nature*, 271.:501.
6. The Exon Theory of Genes. Gilbert, Walter. 1987., *Cold Spring Harb Symp Quant Biol.*, 52.:901-905.
7. The recent origins of introns. Palmer, Jeffrey D. and Logsdon, John M. Jr. 4, 1991. 12, *Current Opinion in Genetics & Development*, 1. : 470-477.
8. Origin of Genes. Gilbert, Walter, Souza, Sandro J. de and Long, Manyuan. 15, 1997. 07 22, *Proceedings of the National Academy of Sciences*, 94.: 7698-7703.
9. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Koonin, Eugene V. 22, 2006. 08 14, *Biology Direct*, 1.
10. Group II twintron: an intron within an intron in a chloroplast cytochrome b-559 gene. Copertino, Donald W.

and Hallick, Richard B. 2, 1991., The EMBO Journal, 10.: 433-442.

11. Spliceosome twin introns in fungal nuclear transcripts. Flipphi, Michel, et al. 2013. 01 19, Fungal Genetics and Biology, 57.:48-57.

12. Alternatively spliced, spliceosomal twin introns in *Helminthosporium solani*. Ág, Norbert, et al. 2015. 10 26, Fungal Genetics and Biology, 85. : 7-13.

13. A mechanism for a single nucleotide intron shift. Fekete, Erzsébet, et al. 15, 2017. 06 08, Nucleic Acids Research, 45. :9085-9092.

14. Emergence and loss of spliceosomal twin introns. Flipphi, Michel, et al. 7, 2017. 10 06, Fungal Biology and Biotechnology, 4.

15. A spliceosomal twin intron (stwintron) participates in both exon skipping and evolutionary exon loss. Kavalecz, Napsugár, et al. 9940, 2019. 07 09, Scientific Reports, 9.

16. Complex intron generation in the yeast genus *Lipomyces*. Ág, Norbert, et al. 6022, 2020. 04 07, Scientific Reports, 10.

17. Internally Symmetrical Stwintrons and Related Canonical Introns in Hypoxylaceae Species. Fekete, Erzsébet, et al. 710, 2021. 08 29, Journal of Fungi, 7.

18. Unique and Repeated Stwintrons (Spliceosomal Twin Introns) in the Hypoxylaceae. Fekete, Erzsébet, et al. 397, 2022. 04 13, Journal of Fungi, 8.

19. Introns and Splicing Elements of Five Diverse Fungi. Kupfer, Doris M., et al. 5, 2004. 10 3, Eukaryotic Cell, 3. : 1088–1100.

20. The Pfam protein families database in 2019. El-Gebali , Sara, et al. D1, 2019., *Nucleic Acids Res*, 47.: D427-D432.
21. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Katoh, Kazutaka and Standley, Daron M. 4, 2013., *Mol Biol Evol*, 30.:772-780.
22. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. Criscuolo, Alexis and Gribaldo, Simonetta. 210, 2010., *BMC Evol Biol*, 10.
23. SMS: Smart Model Selection in PhyML. Lefort , Vincent, Longueville, Jean-Emmanuel and Gascuel, Olivier. 9, 2017., *Mol Biol Evol*, 34.:2422-2424.
24. An improved general amino acid replacement matrix. Le, Si Quang and Gascuel, Olivier. 7, 2008., *Mol Biol Evol*, 25.:1307-1320.
25. Confidence limits on phylogenies: An approach using the bootstrap. Felsenstein, Joseph. 4, 1985., *Evolution*, 39.:783-791.
26. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Anisimova, Maria and Gascuel, Olivier. 4, 2006., *Syst Biol*, 55.:539-552.
27. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Sievers, Fabian, et al. 539, 2011., *Mol Syst Biol*, 7.
28. The Vienna RNA websuite. Gruber , Andreas R., et al. 2008., *Nucleic Acids Res*, 36.:W70-4.

29. Partition function and base pairing probabilities of RNA heterodimers. Bernhart, Stephan H., et al. 3, 2006., Algorithms for Mol Biol, 1.
30. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. Wheeler, Travis J., Clements, Jody and Finn , Robert D. 7, 2014., BMC Bioinform., 15.
31. Mfold web server for nucleic acid folding and hybridization prediction. Zuker, Michael. 13, Nucleic Acids Research : ismeretlen szerző, 2003., 31.:3406–3415.
32. UNAFold: software for nucleic acid folding and hybridization. Markham, Nicholas R. and Zuker, Michael. 2008., Methods Mol Biol, 453.:3-31.
33. RNAalifold: improved consensus structure prediction for RNA alignments. Bernhart, Stephan H., et al. 474, 2008., BMC Bioinformatics, 9.
34. Tempo and mode of genome evolution in the budding yeast subphylum. Shen, Xing-Xing, et al. 175, 2018., Cell, 6. : 1533–1545.
35. Comparative genomics of biotechnologically important yeasts. Riley, Robert, et al. 35, 2016., PNAS, 113.:9882-9887.
36. Characterization of four endophytic fungi as potential consolidated bioprocessing hosts for conversion of lignocellulose into advanced biofuels. Wu, Weihua, et al. 6, 2017., Appl Microbiol Biotechnol, 101.:2603-2618.



Registry number: DEENK/82/2023.PL
Subject: PhD Publication List

Candidate: Fruzsina Szabó-Pénzes

Doctoral School: Pál Juhász-Nagy Doctoral School of Biology and Environmental Sciences

MTMT ID: 10069262

List of publications related to the dissertation

Foreign language scientific articles in international journals (3)

1. Fekete, E., **Pénzes, F.**, Ág, N., Ág-Rácz, V., Karaffa, E. M., Scazzocchio, C., Flipphi, M., Karaffa, L.: Unique and Repeated Stwintrons (Spliceosomal Twin Introns) in the Hypoxylaceae.
J. Fungi. 8 (4), 1-23, 2022. EISSN: 2309-608X.
DOI: <http://dx.doi.org/10.3390/jof8040397>
IF: 5.724 (2021)
2. Fekete, E., **Pénzes, F.**, Ág, N., Scazzocchio, C., Flipphi, M., Karaffa, L.: Internally Symmetrical Stwintrons and Related Canonical Introns in Hypoxylaceae Species.
J. Fungi. 7, 1-21, 2021. EISSN: 2309-608X.
DOI: <http://dx.doi.org/10.3390/jof7090710>
IF: 5.724
3. Ág, N., Kavalecz, N., **Pénzes, F.**, Karaffa, L., Scazzocchio, C., Flipphi, M., Fekete, E.: Complex intron generation in the yeast genus *Lipomyces*.
Sci. Rep. 10, 1-10, 2020. EISSN: 2045-2322.
DOI: <http://dx.doi.org/10.1038/s41598-020-63239-6>
IF: 4.379

Total IF of journals (all publications): 15,827

Total IF of journals (publications related to the dissertation): 15,827

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

16 March, 2023

