Egy új DNS motívum típus in silico jellemzése és szerepe a génszabályozásban
Zárójelentés - OTKA # PD73575, BIOIN

Cserző Miklós

A kutatás első évében az előzetes tervnek megfelelően tudtunk haladni. Kidolgoztunk egy statisztikai módszert a genom vizsgálatára. Modellünk egy fej és egy farok motívum részből áll és a vizsgált genomot a lehetséges fej-farok párok gyakoriságával írjuk le a köztük lévő távtartó hosszának függvényében. Ez az egyes fej-farok mintázatokra kapott frekvencia profilok sorozata. A modellt két változatban is teszteltük: az egyszerűbben a fej és a farok is 5-5 bázisból áll és a távtartó hossza 0 és 54 közé esik, a bonyolultabb modellben a fej és a farok is 6-6 bázis, de 2-2 bázis lötyögés megengedett.

A vizsgálat szempontjából azok a fej-farok mintázatok érdekesek, amelyek frekvencia profilja egy lapos alapvonal képét mutatja egyetlen mély minimummal. Az ilyen profil arra utal, hogy a fej-farok mintázat egy bizonyos kritikus relatív távolságban az átlagosnál lényegesen ritkább. A számítást elvégeztük emberi és éger genomon is. Mindkét esetben nagy számban találtunk ilyen mintázatokat. A mintázatok listája az két genom között ~70%-ban átfed. A talált mintázatokban a direkt fej-farok érintkezés dominál - vagyis a kritikus relatív távolság tipikusan 0.

Az így kapott mintázat-könyvtárak alapján az emberi és egér genomban azonosíthatók azok a szakaszok, ahol a relatív ritka mintázatok koncentráltan fordulnak elő. Elvégezve a számítást azt kaptuk, hogy a mintázat-klaszterek jellemzően 18-25 bázis hosszúak és elsősorban genetikailag kitüntetett helyeken fordulnak elő. A legnagyobb koncentrációban az 5' UTR elején találtunk klasztereket - közvetlenül a transzkripciós starthely mindkét oldalán. Ehhez képest kisebb mértékű felhalmozódást találtunk az exon/intron határok intron felőli oldalán. Harmadsorban a gének 3' UTR szakaszán is találtunk szignifikánsan magas klaszter koncentrációt.

Számos ellenőrző számítást végeztünk különböző módszerekkel előállított referencia szekvenciákkal annak igazolása érdekében, hogy a talált ritka mintázatok nem a már ismert genetikai sajátságok közvetlen következményei, hanem merőben új genetikai jelenségekre utalnak. Ennek során kizártuk a repetitív szekvenciákat és a CpG szigeteket, mint lehetséges okozóit a jelenségnek. Továbbá megállapítottuk, hogy a klasztereink nem mutatnak szignifikáns korrelációt az ismert miRNS szekvenciákkal és az ismert transzkripciós faktor kötőhelyekkel sem esnek egybe.

Viszont jelentős mértékű korrelációt találtunk a kísérletesen meghatározott polimeráz II kötőhelyek és a klasztereink között. Ennél is érdekesebb, hogy a klasztereink nagymértékben átfednek az ENCODE régióban kísérletesen talált nem kódoló kis RNS-ekkel. Ezt az RNS típust egyre több esetben mutatják ki kísérletesen különböző mintákban, viszont senki sem tudja pontosan, hogy mi a szerepük a sejten belül. Abban viszont mindenki egyetért, hogy nem véletlenül vannak jelen. Úgy is

mondhatnánk, hogy a felgyűlt kísérletes adatok mögött jelenleg még hiányzik az elméleti háttér a nem kódoló kis RNS-ek tekintetében. Eddigi eredményeink azt sugalljak, hogy a klasztereink további vizsgálatával közelebb lehet jutni az ncRNS-ek sejten belüli szerepének felderítésében. A talált motívumokat "spanion"-nak neveztük el az ógörög "ritka" szó után.

A kutatás második évében sikerült továbblépnünk az első év eredményeire építve. A kutatási tervnek megfelelően megvizsgáltunk további nyilvánosan elérhető genomokat. A vizsgált genomok: élesztő, neurospora, fonálféreg, gyümölcslégy, malária szúnyog, fugu, zebrahal, tüskés pikó, csirke, kutya, marha, macska, ló és sertés.

Minden vizsgált genomban találtunk 'spanionokat'. Az egyszerűbb élőlények esetén kevesebbet, a fejlettebbek esetén többet. A két vizsgált gomba genomjában éppen csak kimutatható mennyiségben vannak 'spanionok'. Ezek a genomok elég kicsik a többi vizsgálthoz képest és evolúciós értelemben is igen távol állnak tőlük. Ezek a genomok a modellel részletesen nem vizsgálhatók.

A fonálféregtől kezdve viszont a modell egyre növekvő számban azonosított 'spanionokat' az adott élőlény evolúciós fejlettségével arányos mértékben. A csirke genomtól kezdve az összes genom elemzése igen hasonló képet mutat. A 'spanion' listák nagy mértékben átfednek és statisztikus tulajdonságaik is közel azonosak. Így ezek a genomok a modell segítségével jól elemezhetők.

Az emberi genom részletes elemzését megkezdtük a második évben. A 'spanion' klaszterek elsősorban a génátírás kezdőhelye körül fordulnak elő. Közbülső exonokban szinten jelentős mennyiségben vannak 'spanion' klaszterek, de itt nincs egyértelműen kitüntetett genomi elhelyezkedés az exonok végeihez képest. A 3'UTR szakaszok viszonylag kevés 'spanion' klasztert tartalmaznak, viszont ezek a szakaszok a 3’UTR szakaszok 5' végén fordulnak jellemzően elő.

A második kutatási évben nem remélt és igen kedvező fordulat állt be. Taft és munkatársai leközöltek egy kísérletes adatsort, amely humán THP-1 sejtvonalon végzett elemzésüket tartalmazta (VOLUME 41, NUMBER 5, MAY 2009 NATURE GENETICS, pp. 572) Azonosítottak egy 18 - 20 bázis hosszú nem-kódoló RNS sokaságot, amely a génátírás kezdőpontját követő genomi helyről származik és 'tiRNS' néven új RNS típusként közölték. Összevetve az adatbázisukat a statisztikus modell 'spanion' klasztereivel, a két halmaz 70%-ban azonos. Gyakorlatilag a statisztikus modellünk egy igen hatékony tiRNS becslő módszer.

A tiRNS-ek sejten belüli szerepe egyelőre nem ismert, de kézenfekvőnek latszik, hogy a génszabályozás egy teljesen új és általánosan elterjedt formájában vesznek részt. A 'spanion' modellünk igen hasznos lehet a tiRNS-ek kutatásában, mivel lényegében megadják az elméleti alapot a megfigyelések értelmezéséhez.

A humán genomot megvizsgáltuk az észlelt 'spanion' klaszter tartalom szempontjából az egyes gének 5' vége körüli 6 Kb hosszú régióban. A géneket sorba rendezve a 'spanion' klaszter tartalom szerint és kiválasztva az 500, 'spanion' klaszterben leggazdagabb gént azt kaptuk, hogy ezen géneknek megfelelő GO-kifejezésekben statisztikusan szignifikáns mértékben felülreprezentált a génszabályozással kapcsolatos funkció.

A kutatás harmadik évében leközöltuk az első két év eredményeit (Cserzo et al. Biology Direct 2010, 5:56). Közleményünkben ismertetjük a statisztikus elemzés minden részletét az alkalmazott ellenőrző számításokkal együtt. Ismertetjük az ember és egér genomok statisztikus jellemzőit a megtalált

'spanion' klaszterekre vonatkozóan. Leírjuk a 'spanion' klaszterek relativ helyzetét az emberi genomban a kitüntetett genomi pozíciókhoz képest (transzkripciós idító helyek, exon/intorn határok, 'Fig. 2.' az eredeti közleményben). Bemutatjuk a 'spanion' klaszterek egybeesését két, kisérletesen meghatározott adatsorral (RNS polimeráz II kötőhelyek és tiRNS adatbázis, 'Fig. 3 & 4' az eredeti közleményben). Továbbá közöljük a 'spanion' klaszterekben különösen gazdag géneken végzett GO-kifejezés elemzés eredményét ('Fig. 5 & Tab. II' az eredeti közleményben).

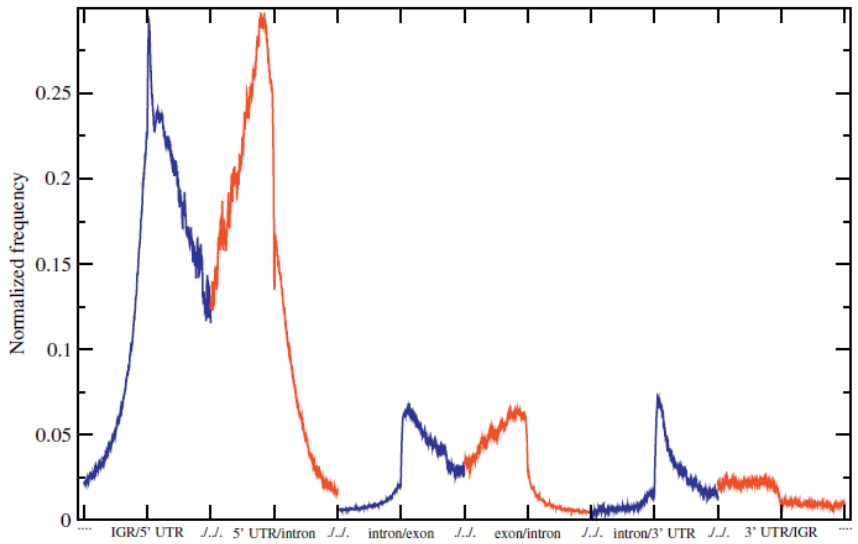

Figure 2 Normalized distribution of *spanion* clusters around distinct genetic landmarks. Human fragments are aligned at six key genetic landmarks and the preferred location of ***spanion*** clusters are presented on the plot. The individual distributions arranged into a single plot concatenating them in head to tail fashion representing a hypothetical gene. The original distributions are represented by alternating colours: ***IGR/5' UTR (blue); 5' UTR/intron (red); intron/exon (blue); exon/intron (red); intron/3' UTR (blue) and 3' UTR/IGR (red).*** In each case the point of the alignment is the junction of the two genetic features in question. The possible number of hits decreases with the distance from the selected landmark as fragments are considered only up to the next genetic borderline or 1 kb distance at most. Therefore the raw counts normalised with the maximal number of possible hits at each given position relative to the aligned landmark.

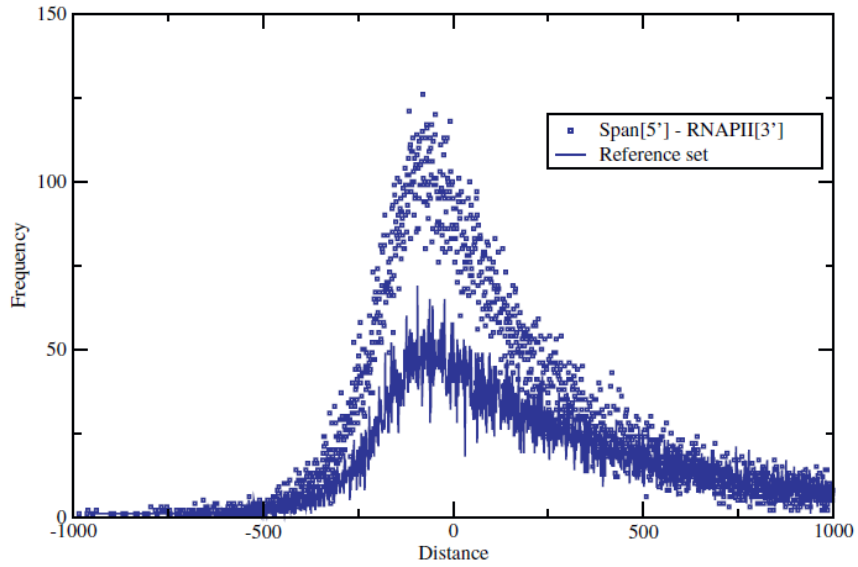

Figure 3 Co-localization of *spanion* clusters and experimental RNA polymerase II sites. The distribution of the distance between the upstream end of a ***spanion*** clusters and the downstream end of the closest reported RNA polymerase II sites in upstream position relative to the ***spanion*** cluster [16]. The original measurement has been done on confluent HEK cells. In the reference dataset the local ***spanion*** cluster patterns of the genes were mapped to the chromosomes using the global coordinates of randomly picked different genes. Real dataset: squares, reference set: continuous line. The distribution is asymmetric due to the way the distance is measured. Alternative metrics were also tested and provided similar results (data not shown). The peak around zero in the reference case reflects to the common general preference of the transcript zero position of the two genetic features. Note that the RNA polymerase II sites span a few hundred bases typically while the most common size of the ***spanion*** clusters is around 25 bases. (See Figure 1 and Figure 1 of Sultan et al [16].)

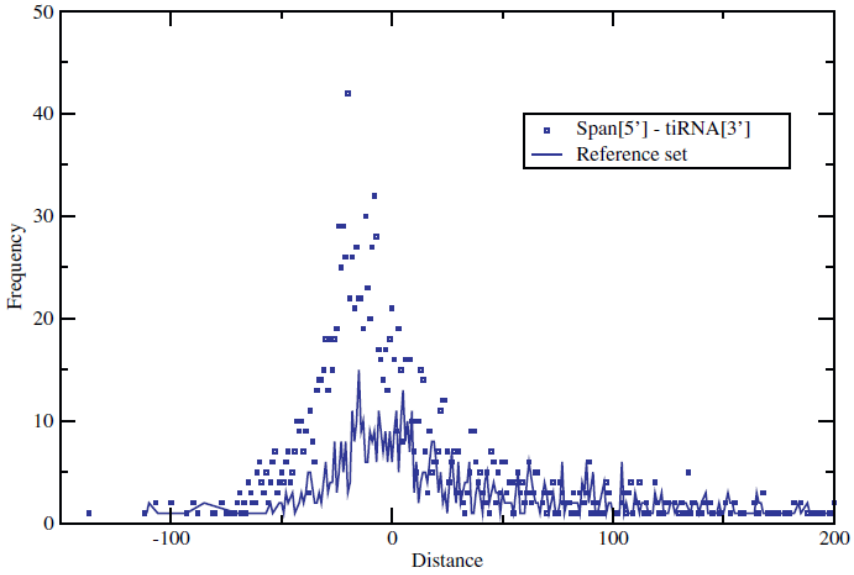

Figure 4 Co-localisation of *spanion* clusters and experimentally detected tiRNAs. The distribution of the distances between the upstream end of a ***spanion*** cluster and the downstream end of the closest reported tiRNA hit in upstream position relative to the ***spanion*** cluster [7]. The real dataset is marked by squares; the reference set (continuous line) was generated as in the case of the RNA polymerase II dataset (see above). The typical size of the tiRNAs and the ***spanion*** clusters is directly comparable: 20 - 25 bases. The peak at -20 indicates that typically the 5' ends of the two kinds of segments are only a couple of bases apart.

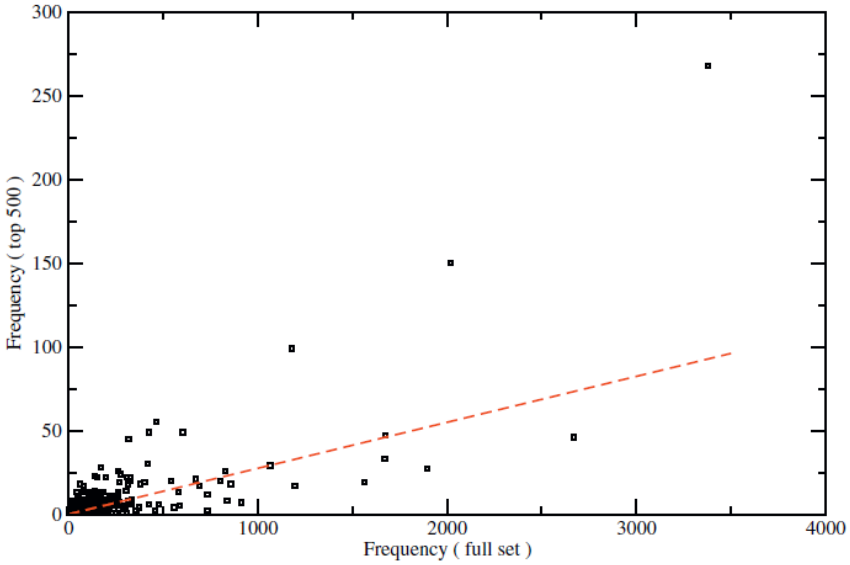

Figure 5 The relation of *spanion* cluster concentration and GO terms. The protein coding Human genes were ranked according to the ***spaninon*** cluster content of the 6 kb transcript proximal region. The frequency of the GO terms associated with the 500 top scoring genes was obtained and compared with the GO term frequencies of the full set. In random case one would expect values proportional with the ratio of the sizes of the two sets. This ratio is indicated by dashed line on the plot. Points above the line correspond to GO terms overrepresented in the short listed set of high ***spanion*** cluster content genes. The 25 most frequent GO terms are listed in Table 2.

Table 2 The 25 most frequent GO terms associated with the 500 genes richest in *spanion* clusters in their 6 kb transcript proximal region

Freq	GO code	GO term
268	GO:0006355	regulation of transcription, DNA-dependent
150	GO:0045449	regulation of transcription
99	GO:0007275	multicellular organismal development
55	GO:0050826	response to freezing
55	GO:0042309	Homoiothermy
49	GO:0045944	positive regulation of transcription from RNA polymerase II promoter
49	GO:0030154	cell differentiation
47	GO:0007186	G-protein coupled receptor protein signaling pathway
46	GO:0007165	signal transduction
45	GO:0000122	negative regulation of transcription from RNA polymerase II promoter
33	GO:0008152	metabolic process
30	GO:0007399	nervous system development
29	GO:0007155	cell adhesion
28	GO:0009887	organ morphogenesis
27	GO:0006468	protein amino acid phosphorylation
26	GO:0007156	homophilic cell adhesion
26	GO:0006412	Translation
24	GO:0006260	DNA replication
23	GO:0007417	central nervous system development
22	GO:0045941	positive regulation of transcription
22	GO:0008284	positive regulation of cell proliferation
22	GO:0007420	brain development
22	GO:0006357	regulation of transcription from RNA polymerase II promoter
21	GO:0007049	cell cycle
20	GO:0016311	Dephosphorylation

Az algoritmus végül is nem bizonyult alkalmasnak internet alapú szolgáltatás beindítására, mert ésszerűtlenül nagy adatforgalmat generált volna. Ehelyett a közlemény függeléke ként leközöltük a ’spanion’ klaszterek teljes listáját emberre és egérre, valamint a ’spanion’ klaszterek becslésére szolgáló program forráskódját is nyilvánossá tettük.

A cikk megjelenése után felvettük a kapcsolatot John Mattick laboratóriumával (University of Queensland) és elkezdtük kiépíteni egy jövőbeni együttműködés alapjait. John Mattick munkatársa, Ryan Taft, érdeklődést mutatott a munkánk iránt és rendelkezésünkre bocsájtott egy lényegesen nagyobb kisérletes kis-RNS adatbázist további elemzés céljából. Ez a munka jelenleg folyik és az előzetes eredményeink igen bíztatóak.

A közles során felfigyeltünk egy érdekes jelenségre. Bár a ’spanion’ klaszterek alulreprezentált motívumokból épülnek fel, ez nem jelenti azt, hogy az így kapott szekvenciák szükségképpen egyediek lennének. A megtalált ’spanion’ klaszterek közt van egyedi, van amelyik néhány példányban fordul elő és van olyan is, amely többszázszor vagy akár néhány ezerszer is előfordul. Ezen az alapon a gének közt felírható egy hálózat, amely összeköti azokat a géneket, amelyekben azonos ’spanion’ klaszterek fordulnak elő, akár direkt vagy komplemeter elrendezésben.

Ezt a megfigyelést megemlítettük a cikkben, de nem fejtettük ki részeltesen. A további kutatást ebbe az irányba tervezzük folyatatni. Ezert együttműködést kezdeményeztünk Hutvágner Györggyel (University of Dundee), aki a kis, nem-kódoló RNS-ek területén nemzetközi szaktekintélynek örvend. Terveink szerint kiválasztunk néhány ’spanion’ klasztert és ezekeket megpróbáljuk kisérletesen kimutatni néhány emberi sejtvonalban. Sikeres előkisérletek után nagy átteresztő képességű

módszerek segítségével megpróbáljuk feltérképezni a teljes kis-RNS spektrumot néhány emberi sejtvonalon. Továbbá célzott kisérletek segítségével megpróbáljuk meghatározni az érintett biológiai jelpályákat és mechanizmusokat, amelyekben a 'spanion' klasztereknek megfelelő RNS-ek részt vehetnek.

A munkát poszter formájában három alkalommal mutattuk be nemzetközi konferenciákon a pályázat futamideje alatt: 1); 'German Conference on Bioinformatics 2009', Halle, Germany, 2); 'CSHL Genome Informatics 2010', Hinxton, UK, 3); 'EMBL|EMBO Symposium: Non-Coding Genome 2010', Heidelberg, Germany. Ez utóbbi különösen eredményes volt, mert itt nyílt alkalom a Ryan Tafttal való személyes találkozóra. A pályázat lejárta előtt a kutatást riport formájában ismertette a 'Project Magazine'. Ez a lap az Európában folyó kutatásokról ad rendszeres áttekintést ezzel támogatva a tudománypolitika valamint az innovatív ipar döntéshozóinak munkáját.