7 Ali Bitkinin Zülal Kodlaşdıran Nüvə Genlərinin Potensial Transkripsiya Start Saytlarının Müəyyənləşdirilməsi

H.F. Quliyeva*, Ə.Ü. Abduləzimova, N.Ş. Mustafayev, İ.Ə. Şahmuradov

*AMEA Molekulyar Biologiya və Biotexnologiya İnstitutu, Mətbuat prospekti, 2A, Bakı AZ1073, Azərbaycan; *E-mail: aliyeva-hokume@mail.ru*

Birləpəli *Oryza sativa* və *Zea mays,* ikiləpəli *Arabidopsis thaliana, Glycine max, Medicago truncatula, Populus trichocarpa* və *Vitis vinifera* bitkilərinin müvafiq surətdə 22258, 23330, 17896, 18226, 17645, 38702 və 11035 zülal kodlaşdıran nüvə genlərinin [1000:+101] 5'-nahiyyələrində (+1: genlərin annotasiya olunmuş start nöqtəsi) TSSPlant kompüter proqramının köməyi ilə potensial transkripsiya start saytlarının (TSS) - promotorların axtarışı həyata keçirilmişdir. Nəticədə 7 bitki növündən 149092 genin hər biri üçün ən azı bir potensial TSS müəyyənləşdirilmişdir. Aşkar olunmuş promotorların həm bütün genlər üzrə, həm də ayrıca olaraq plastid və mitoxondri təyinatlı genlər üzrə müqayisəli təhlili aşkar etmişdir ki, bu orqanizmlərin hamısında bütün hallarda qeyri-TATA promotorlar aşkar üstünlük təşkil edirlər (~30% TATA/70% qeyri-TATA). Hər bir gen üçün annotasiya olunmuş gen başlanğıcına ən yaxın TSS götürülməklə, potensial TSS (TSSp) və gen başlanğıcı arasındakı məsafələrin təhlili göstərmişdir ki, bütün orqanizmlər üzrə genlərin təxminən 70%-i üçün bu məsafə 100 nc-dən çox deyildir. Bu müşahidə TSSPlant proqramının kifayət qədər yüksək axtarış dəqiqliyinə dəlalət edir.

Açar sözlər: *Ali bitki, genom, gen, TATA-boks, promotor, TSS, kompüter analizi*

GİRİŞ

Hazırda, insan, gəmiricilər və meymunlar da daxil olmaqla, genomların oxunması üzrə doqquz yüzə yaxın layihə vardır (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi) və hazırda nukleotid ardıcıllığı məlum olan genomların quruluş və funksiya baxımından annotasiyası nəzəri və praktiki baxımdan ən aktual məsələlərdən biridir.

Genom annotasiyasının mühüm problemlərindən biri promotorların müəyyənləşdirilməsidir. Lakin bu gün genom ardıcıllıqlarında promotorların tapılması sadə məsələ deyildir. Promotor ardıcıllıqlarının gen-səciyyəvi arxitekturası və bu sahədə mövcud biliklərin məhdud olması onların axtarışı üçün ümumi strategiyanın yaradılmasını müşkül problem edir. Genom nahiyyələrinin transkripsiyasını tənzimləyən transkripsiya faktorlarının (TF) qısa (5-15 nukleotid cütü, nc) nukleotid ardıcıllıqlarından ibarət birləşmə saytları (TFBS) əsasən promotorlarda yerləşir. Hər bir promotor belə tənzimləyici elementlərin unikal tərkibi ilə səciyyələnir ki, bu da özünəməxsus gen ekpressiyasını müəyyən edir. Eukariotlarda zülal kodlaşdıran genlərin və həmçinin bəzi kiçik RNT genlərinin transkripsiyası RNT II polimeraza (Pol II) vasitəsilə həyata keçirilir. Transkripsiyanın start saytının (TSS) daxil olduğu Pol II promotor nüvəsi ("*core promoter*) TSS-na (+1) nəzərən -60:+40 rayonunu əhatə edir. Həmin nüvədən əvvəl yerləşən 200-300 nc uzunluğundakı nahiyyə isə proksimal promotordur. Proksimal promotorda transkripsiyanın özünəməxsus qaydada tənzimlənməsi üçün tələb olunan çoxsaylı TFBS-ları yerləşir. Bundan başqa, promotorun proksimal promotordan əvvəldə yerləşən distal hissəsində də TFBS-ları (enhanserlər, saylenserlər, insuleytorlar) vardır (Hebing et al., 2015; Geetu et al., 2014; Tak-Ming et al., 2012; Fred et al., 2016.).

Heyvan və bitki orqanizmlərinin promotorlarının təxminən 30-50%-ində TSS-dan 25-45 nc məsafəsində yerləşən və TATA-boks adlanan element vardır - bu promotorlar TATA-promotor adlanır. Eukariot promotorlarında indiyədək aşkar edilmiş ən konservativ element TATA-boks yüksək dərəcədə ekspressiya olunan genlərin çoxunun promotorunda vardır (Smale et al., 2003; Butler et al., 2002; Lemon et al., 2001; Sandelin et al., 2007; Zuo et al., 2011).

Lakin bir çox gen qruplarında (məsələn, "ev təsərrüfatı genləri; *housekeeping genes*) TATA-boks yoxdur - belə promotorlar qeyri-TATA (*TATA-less*) promotorlar adlanır. Bu tip promotorlarda TSS-nın mövqeyi CpG adaları və/ya "inisiasiya elementi" (*initiator element*), Inr, və/ya DPE (*downstream promoter element*) vasitəsilə tənzimlənir (Suzuki et al., 2001; Cooper et al., 2006; Xu et al., 2016).

İndiyədək tədqiq olunmuş promotorların əksəriyyətində çoxsaylı, alternativ TSS-lər aşkar olunmuşdur (Suzuki et al, 2001; Carninci et al., 2006; Taylor et al. 2006; Davuluri et al., 2008). TSS-ları əksər hallarda gen başlanğıcından əvvəldə yerləşsə də, onlara genlərin "daxilində" - transkripsiya olunan hissələrdə olması da nadir hadisə deyildir (Koch et al, 2008; Shahmuradov et al., 2016).

Hazırda bitkilərin təcrübi yolla müəyyənləşdirilmiş 2 tip promotor (TSS) kolleksiyaları mövcuddur.

1) Tam uzunluqlu komplementar DNT (TUkDNT; *full-length cDNA, FL-cDNA*) ardıcıllıqlarının genom ardıcıllıqları ilə müqayisəsi əsasında müəyyənləşdirilən kolleksiyalar; məsələn, RARGE DB (Sakurai et al., 2005; Akiyama et al., 2014) və ppdb (Yamamoto et al., 2008; Hieno et al., 2014).
2) TSS da daxil olmaqla, promotor dəstləri birbaşa təcrübi yollarla müəyyənləşdirilmiş kolleksiyalar; məsələn, EPD (Dreos et al., 2013, 2015), PlantProm DB (Shahmuradov et al., 2003).

Birinci tip bazalar içərisində ppdb ən böyük məlumat resursudur. Bu bazanın son buraxılışında (versiya 3.0; Hieno et al., 2014; http://ppdb.agr.gifu-u.ac.jp/ppdb/cgi-bin/index.cgi) kəsəkotu (*Arabidopsis thaliana*), düyü (*Oryza sativa*), qovaq (*Populus trichocarpa*) və mamır (*Physcomitrella patens*) bitkilərindən on minlərlə Pol II promotorları üzrə məlumat toplanmışdır. O cümlədən, bu bazada kəsəkotunun zülal kodlaşdıran genlərinin hamısı (27206) və düyünün annotasiya olunmuş 32325 zülal kodlaşdıran genindən 12535 gen üçün TSS məlumatı vardır. Digər tərəfdən, bizim analiz aşkar etmişdir ki, kəsəkotunun və düyünun müvafiq surətdə 7878 (~29%) və 1554 (~13%) geni üçün xəritələşdirilmiş TSS ilə müvafiq genin kodlaşdıran DNT ardıcıllığının (KDA; *coding DNA sequences CDS*) annotasiya olunmuş başlanğıc nöqtəsi arasındakı məsafə, d(TSS,KDA), 10 nc-dən azdır. Məlumdur ki, TSS və KDA başlanğıcı arasındakı nahiyyə - 5'-translyasiya olunmayan rayon (TOR; untranslated region, UTR) ribosomların mRNT ilə birləşməsi translyasiyanı həyat keçirməsi üçün tələb olunur. İndiyədək 5'-TOR-un minimum uzunluğu məlum olmasa da, belə hesab olunur ki, translyasiyanın düzgün və tələb olunan səviyyədə getməsi üçün həmin rayonun uzunluğu, ən azı, 20 nukleotid olmalıdır (Chen et al., 2011; Kim et al., 2014; Hinnebusch et al., 2016). Bu halda belə nəticə hasil olur ki, ppdb bazasında minlərlə gen üçün TSS məlumatları yeni təcrübələrlə dəqiqləşdirilməlidir.

İkinci tip bazalar içərisində ilk buraxılışı 2003-cü ildə təqdim olunmuş PlantProm DB həm tərkibindəki TSS-larının ümumi sayına, həm də toplanmış məlumatların təhlilolunma aspektlərinə görə ən böyük resursdur. Bu bazanın hazırki buraxılışı 86 bitki növündən hər bir TSS üçün ayrılıqda birbaşa təcrübə vasitəsi ilə müəyyənləş-dirilmiş və çap olunmuş məqalə ilə təsbit olunmuş 576 TSS üzrə məlumatlar toplanmışdır (Shahmuradov et al., 2012; http://www.softberry. com/plantprom2016/).

Bu uğurlara baxmayaraq, bitkilərin təcrübi yolla müəyyənləşdirilmiş promotor kolleksiyaları bitki promotorlarının çox kiçik bir hissəsini təşkil edir. Ümumiyyətlə, TSS-larının təcrübi yolla müəyyənləşdirilməsi hələ də bahalı və çətin prosesdir. Bu səbəbdən kompüter analizi vasitəsi ilə TSS-larının aşkar edilməsi səmərəli yanaşma olaraq qalır (Mundade et al., 2014; Suryamohan et al., 2015; Levati et al., 2016).

Son 20 ildə promotorların (TSS-larının) axtarışı üzrə bir sıra kompüter proqramları yaradılmışdır, o cümlədən: EP3 (Abeel et al., 2008), TSSP-TCM (Shahmuradov et al., 2005), TSSP (http://www.softberry.com/berry.phtml?topic=tssp&group=programs&subgroup=promoter), PromPredict (Rangannan et al., 2009; Morey et al., 2011). Bu yaxınlarda bitkilərin Pol II promo-torlarının axtarışı üzrə yeni bir proqram, TSSPlant, yaradılmışdır (Shahmuradov et al., 2017). Həmin proqram bu sahədə indiyədək yaradılmış digər proqramlarla müqayisədə ən yüksək axtarış dəqiqliyi ilə səciyyələnir.

Təqdim olunan işin əsas məqsədi birləpəli və ikiləpəli bitkilərin bəzi nümayəndələrinin (cəmi 7 növün) plastid və mitoxondri təyinatlı zülal kodlaşdıran nüvə genlərinin və onların potensial TSS xəritələrinin kompüter vasitəsi ilə müəyyənləşdirilməsi olmuşdur. Aşağıda həmin araşdırmaların nəticələri verilir və müzakirə olunur.

MATERİAL VƏ METODLAR

Ali bitki genomunda güman edilən promotorların (TSS-larının) axtarışı üçün 7 bitki növünün nüvə genomunun annotasiyasından istifadə edilmişdir (http://plants.ensembl.org/info/website/ftp/index.html): birləpəli düyü (*Oryza sativa*, 35655 gen; genom assembleyası IRGSP-1.0) və qarğıdalı (*Zea mays*; 36988 gen; genom assembleyası AGPv3), ikiləpəli kəsəkotu (*Arabidopsis thaliana*; 27201 gen; genom assembleyası TAIR10), qara yonca (*Medicago truncatula*; 47202 gen; genom assembleyası MedtrA17_4.0), qovaq ağacı (*Populus trichocarpa*; 38449 gen; genom assembleyası JGİ2.0), şərab üzümü (*Vitis vinifera*; 26118 gen; genom assembleyası İGGP_12x) və soya (*Glycine max*; 53151 gen genom assembleyası; v1.0). Analiz üçün, getseqPP kompüter proqramından (Shahmuradov, çap olunmamışdır) istifadə etməklə, annotasiya olunmuş zülal ardıcıllıqları və müvafiq genlərin [-1000: +101] rayonları (+1: genin start nöqtəsi) götürülmüşdür. Promotor axtarışı üçün yalnız uzunluğu 20 nc və daha çox olan 5'-TOR ilə annotasiya olunmuş zülal kodlaşdıran genlər seçilmişdir; bir neçə başlanğıc nöqtəsi annotasiya olunmuş genlər üçün ən uzun 5'-TOR-a uyğun başlanğıc nöqtəsi götürülmüşdür.

Genomda gen duplikasiyaları nəticəsində bir neçə nüsxə ilə təmsil olunan genlərdən və müvafiq zülallardan yalnız biri seçilmişdir. Bu məqsədlə əvvəlcə zülalların və müvafiq genlərin promotor nahiyyələrinin BLAST proqram paketi (Altschul et al., 2004) vasitəsi ilə növdaxili cüt-cüt müqayisəsi aparılmış və daha sonra BLAST nəticələri BLAN və getseqPP (Şahmuradov, çap olunmamışdır) kompüter proqramlarının köməyi ilə təhlil edilmiş, tam uzunluqlu və oxşarlıq dərəcəsi 90%-dən aşağı olmayan ardıcıllıqlardan yalnız biri götürülmüşdür. Nəticədə, *O. sativa*, *Z. mays*, *A. thaliana*, *M. truncatula*, *P. trichocarpa*, *V. vinifera* və *G. max* bitkilərindən müvafiq surətdə 22258, 23330, 17896, 18226, 17645, 11035 və 38702 zülal və promotor ardıcıllıqları seçilmişdir.

TSS-larının axtarışı Şahmuradov, Umarov və Solovyov tərəfindən bu yaxınlarda yaradılmış TSSPlant kompüter proqramı (www.cbrc.kaust.edu.sa/download/files/TSSPlant_linux.tar.gz) vasitəsi ilə aparılmışdır.

Zülalların mümkün təyinat yerləri ProtComp proqramının vasitəsi ilə müəyyənləşdirilmişdir (http://www.softberry.com/berry.phtml?topic=protcomppl&group=programs&subgroup=proloc).

NƏTİCƏLƏR VƏ ONLARIN MÜZAKİRƏSİ

Plastid və mitoxondri təyinatlı genlərin proqnozlaşdırılması. *O. sativa*, *Z. mays*, *A. thaliana*, *M. truncatula*, *P. trichocarpa*, *V. vinifera* və *G. max* bitkilərindən 149092 zülal ardıcıllığı ProtComp kompüter proqramı vasitəsi ilə analiz edilmişdir və hər növ üçün potensial plastid və mitoxondri təyinatlı zülal ardıcıllıqları dəstləri müəyyənləşdirilmişdir. Kəsəkotunda 3997 plastid və 1533 mitoxondri; soyada 5256 plastid və 2339 mitoxondri; qara yoncada 3759 plastid və 1569 mitoxondri, qovaq ağacında 3215 plastid və 1307 mitoxondri, şərab üzümündə 2093 platid və 832 mitoxondri, düyüdə 4525 plastid və 1761 mitoxondri, qarğıdalıda 5159 plastid və 1933 mitoxondri təyinatlı nüvə genləri aşkar edilmişdir (Cədvəl 1). Bu analizin maraqlı nəticələrindən biri tranzit peptidlərinə görə plastidlərə yaxud mitoxondrilərə ünvanlanması güman edilən zülalların ümumi sayı üzrə növlərarası böyük fərq müşahidə olunur. Çox güman ki, müşahidə olunan fərq müxtəlif növlərdə annotasiya olunmuş genlərin (zülalların) sayında olan fərqlə bağlıdır. Lakin həmin fərq hər bir növün özünəməxsus xüsusiyyətləri ilə də bağlı ola bilər. Hazırda bu istiqamətdə araşdırmalarımız davam edir.

Potensial transkripsiya start saytlarının müəyyənləşdirilməsi. TSSPlant proqramının köməyi ilə 7 bitki növündən 149092 genin hər birinin 5'-nahiyyəsində ən azı bir potensial promotor (TSS) aşkar edilmişdir. Aşkar olunmuş promotorlar həm bütün genlər üzrə, həm də ayrıca olaraq plastid və mitoxondri təyinatlı genlər üzrə müqayisəli təhlil olunmuşdur. Alınmış nəticələrin əsas inteqral məqamları Cədvəl 1-də verilmişdir. Bu analizin məqamları

Cədvəl 1. 7 ali bitkinin zülal kodlaşdıran nüvə genlərinin potensial transkripsiya start saytları üzrə ümumi statistik göstəricilər

Orqanizm	Zülal genləri dəsti	Dəstdəki genlərin sayı	TATA promotorlar	Qeyri-TATA promotorlar	TATA/qeyri-TATA, %
A.t.	Bütün genlər	17896	5534	12362	30.9/69.1
	Plastid təyinatlı	3994	1266	2728	31.7/68.3
	Mitoxondri təyinatlı	1532	469	1063	30.6/69.4
G.m.	Bütün genlər	38702	5833	32896	15.1/84.9
	Plastid təyinatlı	5252	669	4583	12.7/87.3
	Mitoxondri təyinatlı	2339	332	2007	14.2/85.8
M.t.	Bütün genlər	18226	3600	14626	19.8/80.2
	Plastid təyinatlı	3759	657	3102	17.5/82.5
	Mitoxondri təyinatlı	1569	268	1301	17.1/82.9
P.t.	Bütün genlər	17645	4016	13629	22.8/77.2
	Plastid təyinatlı	3215	623	2592	19.4/80.6
	Mitoxondri təyinatlı	1307	264	1043	20.2/79.8
V.v.	Bütün genlər	11035	2999	8036	27.2/72.8
	Plastid təyinatlı	2083	467	1616	22.4/77.6
	Mitoxondri təyinatlı	829	201	628	24.2/75.8
O.s.	Bütün genlər	22258	7457	14801	33.5/66.5
	Plastid təyinatlı	4508	1310	3198	29.1/70.9
	Mitoxondri təyinatlı	1756	519	1237	29.6/70.4
Z.m.	Bütün genlər	23330	8949	14381	38.4/61.6
	Plastid təyinatlı	5137	1825	3312	35.5/64.5
	Mitoxondri təyinatlı	1921	681	1240	35.5/64.5

A.t. – Arabidopsis thaliana; M.t. – Medicago truncatula; P.t. – Populus trichocarpa; G.m. – Glycine max; V.v. – Vitis vinifera; O.s. – Oryza sativa; Z.m. – Zea mays.

Cədvəl 1-də verilmişdir. Bu analizin aşkar etdiyi ümumi və maraqlı fakt odur ki, tədqiq olunmuş orqanizmlərin hamısında həm bütün genlər, həm də yalnız plastid yaxud mitoxondri təyinatlı genlərdə qeyri-TATA tipli promotorlar aşkar üstünlük təşkil edirlər.

Daha sonra, hər bir orqanizm üçün ayrılıqda həm bütün genlər, həm də yalnız plastid və mitoxondri təyinatlı genlər üzrə aşkar edilmiş potensial TSS-ları ilə müvafiq genlərin annotasiya olunmuş gen başlanğıcı arasındakı məsafələr hər iki promotor sinfi (TATA və qeyri-TATA) üzrə birlikdə və ayrı-ayrılıqda təhlil olunmuşdur (**Cədvəl 2, 3, 4**). Əlavə olaraq, nümunə kimi, şəkil 1-də bütün genlər üzrə qeyd olunan məsafələrin paylanma histoqramı da verilir.

Hər bir gen üçün annotasiya olunmuş gen başlanğıcına ən yaxın TSS götürülməklə, TSSPlant vasitəsi ilə tapılmış potensial TSS (TSSp) və gen başlanğıcı arasındakı məsafə hesablanmışdır: analiz olunmuş genlərin ~51,2%-ində potensial TSS annotasiya olunmuş gen başlanğıcının yaxınlığında (≤50 nc məsafədə) yerləşir; TSSp və gen başlanğıcı arasında məsfafə 100 nc-dən çox olmayan genlər isə bütün genlərin ~70%-ini təşkil edir. O cümlədən, genlərin annotasiya olunmuş başlanğıcı ətrafında (≤50 nc məsafədə):

- bütün genlər üzrə, kəsəkotunda 10355, soyada 20202, qara yoncada 10477, qovaq ağacında 8222, şərab üzümündə 4970, düyüdə 11827 və qarğıdalıda 11993 TSS aşkar olunmuşdur;
- plastid təyinatlı genlər üzrə, kəsəkotunda 2268, soyada 2701, qara yoncada 2197, qovaq ağacında 1444, şərab üzümündə 928, düyüdə 2411 və qarğıdalıda 2658 TSS aşkar olunmuşdur;
- mitoxondri təyinatlı genlərdə kəsəkotunda 871, soyada 1183, qara yoncada 898, qovaq ağacında 586, şərab üzümündə 376, düyüdə 936 və qarğıdalıda 986 TSS aşkar edilmişdir.

Bu nəticələr göstərir ki, TSSPlant proqramı vasitəsi ilə müəyyənləşdirilmiş potensial TSS-ları əksər hallarda annotasiya olunmuş gen başlanğıcı ətrafında yerləşir. Potensial TSS-larının annotasiya başlanğıcından nisbətən uzaqda yerləşdiyi hallar ilə bağlı qeyd etmək lazımdır ki, istifadə olunmuş genom annotasiyalarında genləri başlanğıcı mRNT (transkriptom) məlumatlarının məhdudluğu üzündən bir çox hallarda dəqiq deyildir.

Cədvəl 2. 7 ali bitkinin zülal kodlaşdıran nüvə genlərinin annotasiya olunmuş başlanğıc nöqtəsi ilə həmin başlanğıca ən yaxın potensial TSS arasındakı məsafə üzrə ümumi statistik göstəricilər (bütün genlər üçün)

Orqanizm	Promotor sinfi	Potensial TSS və annotasiya olunmuş genin başlanğıcı arasındakı məsafə, nc					
		0-50	51-100	101-200	201-400	401-600	> 600
A.t.	cəmi	10355	2447	3474	1583	28	9
	TATA	3489	688	907	421	22	7
	qeyri-TATA	6866	1759	2567	1162	6	2
G.m	cəmi	20202	5470	8902	4075	40	13
	TATA	4084	596	787	347	18	1
	qeyri-TATA	16118	4874	8115	3728	22	12
M.t	cəmi	10477	2361	3610	1760	15	3
	TATA	2578	377	461	175	8	1
	qeyri-TATA	7899	1984	3149	1585	7	2
O.s.	cəmi	11827	2745	4351	2853	309	173
	TATA	3986	759	1436	1070	139	67
	qeyri-TATA	7841	1986	2915	1783	170	106
P.t	cəmi	8222	2447	4526	2392	50	8
	TATA	2632	433	609	321	17	4
	qeyri-TATA	5590	2014	3917	2071	32	4
V.v.	cəmi	4970	1497	2654	1688	148	78
	TATA	1724	359	487	316	76	37
	qeyri-TATA	3246	1138	2167	1372	72	41
Z.m	cəmi	11993	2713	4751	3169	483	221
	TATA	4428	937	1894	1357	229	104
	qeyri-TATA	18995	1776	2857	1812	254	117

A.t. – Arabidopsis thaliana; M.t. – Medicago truncatula; P.t. – Populus trichocarpa; G.m. – Glycine max; V.v. – Vitis vinifera; O.s. – Oryza sativa; Z.m. – Zea mays.

Cədvəl 3. 7 ali bitkinin zülal kodlaşdıran nüvə genlərinin annotasiya olunmuş başlanğıc nöqtəsi ilə həmin başlanğıca ən yaxın potensial TSS arasındakı məsafə üzrə ümumi statistik göstəricilər (yalnız plastid təyinatlı genlər üçün)

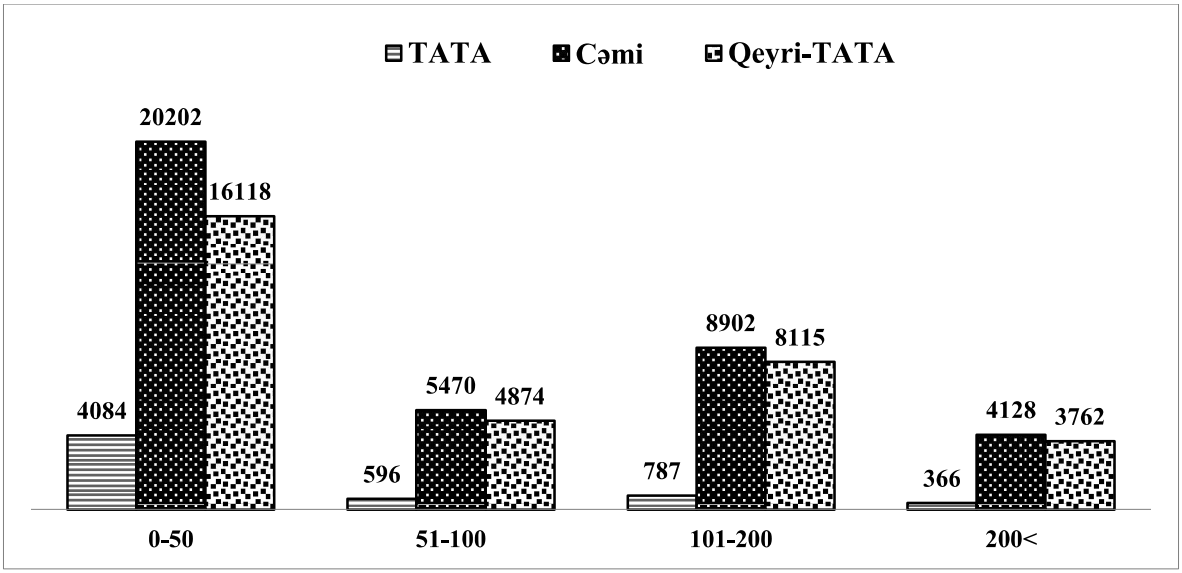
Orqanizm	Promotor sinfi	Potensial TSS və annotasiya olunmuş genin başlanğıcı arasındakı məsafə, nc					
		0-50	51-100	101-200	201-400	401-600	> 600
A.t.	cəmi	2268	592	765	360	5	4
	TATA	755	187	219	98	3	4
	qeyri-TATA	1513	405	546	262	2	0
G.m	cəmi	2701	734	1267	546	3	1
	TATA	431	79	110	46	3	0
	qeyri-TATA	2270	655	1157	500	0	1
M.t	cəmi	2197	486	727	343	6	0
	TATA	442	85	84	43	3	0
	qeyri-TATA	1755	401	643	300	3	0
O.s.	cəmi	2411	536	899	584	54	24
	TATA	660	123	283	209	27	8
	qeyri-TATA	1751	413	616	375	27	16
P.t	cəmi	1444	483	856	417	12	3
	TATA	401	80	101	36	3	2
	qeyri-TATA	1043	403	755	381	9	1
V.v.	cəmi	928	290	503	325	13	14
	TATA	261	58	74	57	8	9
	qeyri-TATA	667	232	429	268	15	5
Z.m	cəmi	2658	596	1067	701	72	43
	TATA	844	221	415	296	68	19
	qeyri-TATA	1814	375	652	405	42	24

A.t. – Arabidopsis thaliana; M.t. – Medicago truncatula; P.t. – Populus trichocarpa; G.m. – Glycine max; V.v. – Vitis vinifera; O.s. – Oryza sativa; Z.m. – Zea mays.

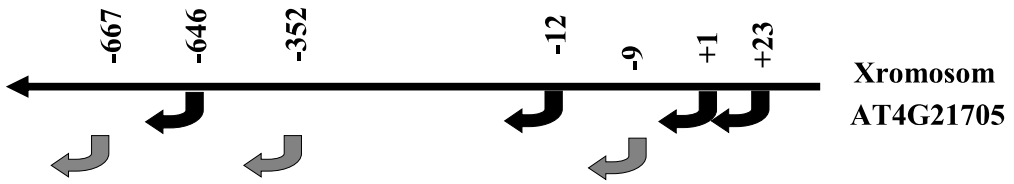
Cədvəl 4. 7 ali bitkinin zülal kodlaşdıran nüvə genlərinin annotasiya olunmuş başlanğıc nöqtəsi ilə həmin başlanğıca ən yaxın potensial TSS arasındakı məsafə üzrə ümumi statistik göstəricilər (yalnız mitoxondri təyinatlı genlər üçün)

Orqanizm	Promotor sinfi	Potensial TSS və annotasiya olunmuş genin başlanğıcı arasındakı məsafə, nc					
		0-50	51-100	101-200	201-400	401-600	> 600
A.t.	cəmi	871	215	296	146	4	0
	TATA	277	64	84	40	4	0
	qeyri-TATA	594	151	212	106	0	0
G.m	cəmi	1183	361	529	262	3	1
	TATA	219	33	53	25	2	0
	qeyri-TATA	964	328	476	237	1	1
M.t	cəmi	898	208	305	157	1	0
	TATA	184	25	38	21	0	0
	qeyri-TATA	714	183	267	136	2	0
O.s.	cəmi	936	196	346	236	33	9
	TATA	247	50	109	92	17	4
	qeyri-TATA	689	146	237	144	16	5
P.t	cəmi	586	212	344	163	2	0
	TATA	174	32	42	16	0	0
	qeyri-TATA	412	80	302	147	2	0
V.v.	cəmi	376	116	179	139	15	4
	TATA	115	19	32	26	8	1
	qeyri-TATA	261	97	147	113	7	3
Z.m	cəmi	986	251	385	251	30	18
	TATA	309	87	160	105	12	8
	qeyri-TATA	677	164	225	146	18	10

A.t. – Arabidopsis thaliana; M.t. – Medicago truncatula; P.t. – Populus trichocarpa; G.m. – Glycine max; V.v. – Vitis vinifera; O.s. – Oryza sativa; Z.m. – Zea mays.



Şəkil 1. Soya bitkisinin annotasiya olunmuş bütün nüvə genlərinin başlanğıcı ilə ona ən yaxın güman olunan TSS arasındakı məsafələrin TATA və qeyri-TATA sinifləri üzrə ayrı-ayrılıqda və birlikdə paylanması.



+23 (11534824) məlum TSS	**naməlum**
+1 annotasiya olunmuş gen başlanğıcı	
-9 (11534856) güman olunan TSS	**TATA-**
-12 (11534859) məlum TSS	**naməlum**
-352 (11535199) güman olunnan TSS	**TATA+**
-646 (11535493) məlum TSS	**naməlum**
-667 (11535514) güman olunan TSS	**TATA+**

Şəkil 2. Kəsəkotu bitkisinin tərkibində pentatrikopeptid təkrarları olan zülalı kodlaşdıran AT4G21705 geninin məlum və güman olunam TSS-larının annotasiya olunmuş gen başlanğıcına və bir-birinə nəzərən yerləşməsinin sxemi. TSS-larının xromosom koordinatları TAIR *A.thaliana* ver6.0 genom annotasiyasına uyğundur. "Naməlum": promotor sinfi məlum deyildir.

Əksər genlər üçün birdən çox potensial TSS-nın tapılması faktı əksər genlərin transkripsiyasının çoxsaylı alternativ promotorlardan (TSS-larından) həyata keçirilməsi ilə bağlı təcrübi faktlarla uzlaşır (Davuluri et al., 2008). Məsələn, kəsəkotu bitkisinin tərkibində pentatrikopeptid təkrarları olan zülalı kodlaşdıran AT4G21705 genində təcrübi yolla 4 TSS müəyyənləşdirilmişdir (Lurin et al., 2004). Bizim TSSPlant analizimiz isə həmin genin promotor nahiyyəsində 3 potensial TSS aşkar etmişdir. Həmin təcrübi və potensial TSS-larının bir-birinə nəzərən yerləşməsinin müqayisəsi (şəkil 2) TSSPlant proqramının axtarış dəqiqiliyi yüksəkdir.

MİNNƏTDARLIQ

Bu iş Azərbaycan Respublikasının Prezidenti yanında Elmin İnkişafı Fondunun maliyyə yardımı ilə yerinə yetirilmişdir – Qrant № EIF/GAM-3-2014-6(21)-24/16/3.

ƏDƏBİYYAT

Abeel T., Saeys Y., Bonnet E., Rouze P., van de Peer Y. (2008) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, **18:** 310-23.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSIBLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **25:** 3389-3402.

Akiyama K., Kurotani A., Iida K., Kuromori T., Shinozaki K., Sakurai T. (2014) RARGE II: an integrated phenotype database of Arabidopsis mu tant traits using a controlled vocabulary. *Plant Cell Physiol;* **55(1):** e4.

Butler J.E.Ş Kadonaga J.T. (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.,* **16:** 2583–2592.

Cooper S.J. Trinklein N.D., Anton E.D. et al. (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.*, **16:** 1-10.

Carninci P., Sandelin A., Lenhard B. et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.,* **38:** 626-635.

Chen Ch., Lin Hy., Pan C.L., Chen F. C. (2011) The genomic features that affect the lengths of 5' untraslated regions in multicellular eukariotes. *BMC Bioinformatics.* **12 Suppl 9:** S3.

Davuluri R.V., Suzuki Y., Sugano S., Plass C., Huang T.H. (2008) The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet.*, **24:** 167-177.

Dreos, R., Ambrosini, G., Périer, R., Bucher, P. (2013) EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucl. Acids Res.,* **41:** D157-D164.

Dreos R., Ambrosini G., Périer, R., Bucher, P. (2015) The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucl. Acids Res.* **43:** D92-D96.

Fred Y.P., Zhiqiu H., Rong-Cai Y. (2016). Bioinformatic prediction of transcription factor binding sites at promoter regions of genes for photoperiod and vernalization responses in model and temperate cereal plants. *BMC Genomics.* **17:** 573.

Geetu T., Karen Betancourt M., Tisha C., Jenny C., Aaron M.W., Gill B. (2014) Automated discovery of tissue-targeting enhancers and transcription factors from binding motif and gene function data. *PLoS Comput Biol.*, **10(1):** e1003449.

Hieno A., Naznin H.A., Hyakumachi M., Sakurai T., Tokizawa M., Koyama H. et al. (2014) ppdb: plant promoter database version 3.0. *Nucl. Acids Res.*, **42:** D1188–D1192.

Hebing C., Hao L., Feng L., Xiaofei Z., Shengqi W., Xiaochen B., Wenjie S. (2015) An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape. *Sci. Rep.*, **5:** 8465.

Hinnebusch A.G., Ivanov I., Sonenberg N. (2016) Translational control by 5'-untranslated regions of eukariotic mRNAs. *Sciense* **352(6292):** 1413-1416.

Kim Y., Goeun L., Eunhyun J. et al. (2014) The immadiate upstream region of the 5'-UTR from the AUG start codon has a pronounced effect on the trabslational efficiency in *Arabidopsis thaliana. Nucleic acid Res.* **42(1):** 485-498.

Koch F., Jourquin F., Ferrier P., Andrau J-C. (2008) Genomewide RNA polymerase II: not genes only! *TIBS,* **33:** 265-273.

Lemon B., Tjian R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.,* **14:** 2551-2569.

Levati E., Sartini S., Ottonello S., Montanini B. (2016) Dry and wet approaches for genome-wide functional annotation of conventional and unconventional transcriptional activators. *Comput. Struct. Biotechnol. J.*, **14:** 262–270.

Lurin C., Andrés C., Aubourg S., Bellaoui M., Bitton F., Bruyère C., Caboche M., Debast C., Gualberto J., Hoffmann B. et al. (2004) Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell.,* **16:** 2089–2103.

Mundade R., Ozer H.G., Wei H., Prabhu L., Lu T. (2014) Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle*, **13:** 2847-2852.

Morey C., Mookherjee S., Rajasekaran G., Bansal M. (2011) DNA free energy-based promoter prediction and comparative analysis of *Arabidopsis* and rice genomes. *Plant Physiol.*, **156:** 1300-1315.

Rangannan V., Bansal M. (2009) Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition. *Mol. Biosyst.*, **5:** 1758–1769.

Sakurai T., Satou M., Akiyama K., Iida K., Seki M., Kuromori T., Ito T., Konagaya A., Toyoda T., Shinozaki K. (2005) RARGE: a large-scale database of RIKEN Arabidopsis resources ranging from transcriptome to phenome. *Nucleic Acids Res.* **33:** D647-650.

Shahmuradov I.A., Gammerman A.J., Hancock J.M., Bramley P.M., Solovyev V.V. (2003) Plant Prom: a database of plant promoter sequences. *Nucl. Acids. Res.*, **31:** 114-117.

Shahmuradov I.A., Solovyev V.V. (2015) Nsite, NsiteH and NsiteM computer tools for studying

transcription regulatory elements. *Bioinformatics*, **31:** 3544-3545.

Shahmuradov I.A., Solovyev, V.V., Gammerman A.J. (2005) Plant promoter prediction with confidence estimation. *Nucl. Acids Res.*, **33:** 1069-1076.

Shahmuradov I.A., Abdulazimova A.U., Khan F.Z., Solovyev V.V., Mustafayev N.Sh., Akbarova Y.Yu., Qamar R., Aliyev J.A. (2012) The PlantProm DB: recent updates. In: *Proceedings of the 2012 International Conference on Biomedical Engineering and Biotechnology* (iCBEB), China, Macau, 612-614.

Shahmuradov I.A., Mohammad R.R., Bougouffa A., Radovanovich A., Bajic V.B. (2017) bTSSfinder: a novel tool for the prediction of promoters in Cyanobacteria and *Escherichia coli. Bioinformatics,* **33 (3):** 334-340.

Shahmuradov I.A., Umarov R.K., Solovyev V.V. (2017) TSSPlant: a new tool for prediction of plant Pol II promoters. *Nucl Acids Res*., **doi: 10.1093/nar/gkw1353.**

Smale S.T., Kadonaga J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem*., **72:** 449-479.

Sandelin A. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews*, **8:** 424-436

Suzuki Y. et al. (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.,* **11:** 677-684.

Suryamohan K., Halfon M.S. (2015) Identifying transcriptional cis-regulatory modules in animal genomes. *Wiley Interdiscip. Rev. Dev. Biol.*, **4:** 59-84.

Taylor M.S. (2006) Heterotachy in mammalian promoter evolution. *PLoS Genet.*, **2:** e30

Tak-Ming C., Kwong-Sak L., Kin-Hong L., Man-Hon W., Terrence C.K.L., Stephen K.W.T. (2012) Subtypes of associated protein–DNA (Transcription Factor-Transcription Factor Binding Site) patterns. *Nucleic Acids Res.*, **40(19):** 9392-9403.

Xu M., Gonzalez-Hurtado E., Martinez E. (2016) Core promoter-specific gene regulation: TATA box selectivity and Initiator-dependent bi-directionality of serum response factor-activated transcription. *Biochhim. Biophys. Acta*, **1859(4):** 553-563.

Yamamoto Y.Y., Obokata J. (2008) ppdb: a plant promoter database. *Nucl Acids Res.*, **33:** D977-D981.

Zuo Y.C., Li Q.Z. (2011) Identification of TATA and TATA-less promoters in plant genomes by intefrating diversity measure, GC-skew and DNA geometric flexibility. *Genomics*, **97:** 112-120.

Определение Потенциальных Сайтов Старта Транскрипции Белок-кодирующих Ядерных Генов в 7-и Высших Растениях

Х.Ф. Кулиева, А.У. Абдулазимова, Н.Ш. Мустафаев, И.А. Шахмурадов

Институт молекулярной биологии и биотехнологий НАН Азербайджана

С помощью компьютерной программы TSSPlant проведен поиск возможных сайтов старта транскрипции (ССТ) – промоторов в [-1000:+101] районах (+1: аннотированное начало гена) 22,258, 23,330, 17,896, 18,226, 17,645, 38,702 и 11,035 (всего 149,092) белок-кодирующих генов, соответственно у *Oryza sativa* и *Zea mays* (однодольные*)*, *Arabidopsis thaliana, Glycine max, Medicago truncatula, Populus trichocarpa* (двудольные) и *Vitis vinifera.* Для каждого гена найден, как минимум, один СС. Сравнительный анализ этих ССТ по классу промоторов для всех генов, в том числе, для пластидных и митохондриальных генов в отдельности, выявил, что у всех растений подавляющее большинство промоторов относится к не-TATA промоторам (~70% не-TATA промоторы против ~30% TATA промоторы). Анализ расстояний между потенциальным TSSp и началом гена для всех изученных видов, показал, что при условии использования самого близкого к аннотированному началу гена ССТ, для 70% и больше генов это расстояние составляет менее 100 нуклеотидов. Последнее наблюдение указывает на достаточно высокую предсказательную точность программы TSSPlant.

Ключевые слова: *Высшее растение, геном, ген, TATA-бокс, промотор, ССТ, компьютерный анализ*

Identification Potential Transcription Start Sites of Protein Encoded Nuclear Genes in 7 Higher Plants

H.F. Guliyeva, A.U. Abdulazimova, N.Sh. Mustafayev, I.A. Shahmuradov

Institute of Molecular Biology and Biotechnologies, Azerbaijan National Academy of Sciences

Using the computer program TSSPlant, search for putative transcription start sites (TSS) – promoters in [-1000:+101] regions (+1 is the annotated gene start) of 22,258, 23,330, 17,896, 18,226, 17,645, 38,702 and 11,035 (totally, 149,092) protein-coding genes from monocot *Oryza sativa* and *Zea mays,* dicot *Arabidopsis thaliana, Glycine max, Medicago truncatula, Populus trichocarpa* and *Vitis vinifera*, respectively, was performed. At least, one potential TSS for every gene was predicted. The comparative analysis of these TSSs by the promoter class for all genes, as well as for only plastid or mitochondrial genes revealed that in all plants TATA-less promoters prevail over the TATA-promoters (~70% TATA-less promoters vs ~30% TATA-promoters). Taking, for every gene, only the predicted TSS (TSSp) which is located closest to the annotated gene start, an analysis of distances between TSSp and gene starts showed that for 70% and more genes this distance is less than 100 bp. These findings indicate that the prediction accuracy of TSSPlant program is quite high.

Keywords: *Higher plant, genome, gene, TATA-box, promoter, TSS, computational analysis*