Gerinces és növényi ortológ promóter adatbázisok fejlesztése és elemzése

Doktori (Ph.D.) értekezés

Készítette: Sebestyén Endre

Eötvös Loránd Tudományegyetem

Természettudományi Kar Biológia Doktori Iskola

Vezetője: Dr. Erdei Anna, egyetemi tanár, az MTA levelező tagja

Klasszikus és Molekuláris Genetika Doktori Program

Vezetője: Dr. Orosz László, egyetemi tanár, az MTA rendes tagja

Témavezető: Dr. Barta Endre, tudományos főmunkatárs

Kutatóhely: Mezőgazdasági Biotechnológiai Kutatóközpont, Gödöllő

Budapest, 2011

Tartalom

1. Bevezetés

1.1. Az eukarióta promóter

Az eukarióta promóterek és a transzkripció szabályozásának elemzése és megértése az elmúlt évek-évtizedek egyik központi témája lett a genetika és genomika területén. Az egyre nagyobb számú teljes genomszekvenciának, az új generációs szekvenáló technikáknak, a különféle array alapú kísérleteknek, az egyre kifinomultabb bioinformatikai módszereknek és a növekvő számítási kapacitásnak köszönhetően nagyszámú szabályozó régió, azaz promóter vizsgálatára van lehetőség. Mivel az eukarióta promóterek nagyságrendekkel összetettebben működnek, mint a kezdetben elemzett prokarióta promóterek, vizsgálatukhoz és a különféle biológiai mechanizmusok megértéséhez is teljesen új és más alapokra épülő módszerek szükségesek, mint kezdetben.

Az eleinte általánosnak vélt promóter elemek és működési mechanizmusok, mint például a TATA-box, úgy tűnik nem találhatók meg minden gén promóterében, és az azonos vagy hasonló expressziós mintázatot mutató, esetleg ortológ géncsoportok esetében is sokkal nehezebb a feltételezhetően biológiai szereppel bíró transzkripciós faktor kötőhelyek felkutatása, és azok alapján a gének kifejeződésének, pontos működésének előrejelzése.

1.2. Transzkripciós szabályozás 'in silico' vizsgálata

A promóterek és a transzkripciós szabályozás vizsgálatában nagy segítséget nyújthatnak a különböző *in silico* bioinformatikai módszerek. A promóterek és a különböző ismert vagy ismeretlen transzkripciós faktor kötőhelyek elemzésére ma már több tucat algoritmus, statisztikai módszer, program vagy teljes programcsomag áll rendelkezésre. A módszereket két nagy csoportra különíthetjük: vagy a kísérletes eredményekre támaszkodunk, és az eddigiekben már leírt transzkripciós faktor kötőhelyeket keresünk, vagy *de novo*, még le nem írt kötőhelyeket próbálunk meg felderíteni különböző statisztikai és/vagy összehasonlító genomikai, bioinformatikai módszerekkel.

Az ismeretlen kötőhelyek előrejelzése esetén a másik alapvető választóvonal a módszerek között az, hogy felhasználjuk-e valamilyen formában a szekvenciákhoz kapcsolható filogenetikai információt. Abban az esetben, ha egy adott fajon belül a hasonló időpontban, szövetben vagy szervben, esetleg hasonló környezeti hatásra kifejeződő gének egy csoportjának a promótereit vizsgáljuk, nincs felhasználható filogenetikai információ az elemzés során. Ha a promóterek különböző fajok azonosan vagy hasonlóan kifejeződő génjeiből vagy géncsoportjaiból származnak, akkor a filogenetikai információ, azaz a szekvenciák közötti evolúciós konzerváltság mértéke és a konzerválódott régiók elhelyezkedése ötleteket adhat a vélt funkcionális kötőhelyek elhelyezkedését illetően.

2. Irodalmi áttekintés

2.1. Az eukarióta promóter

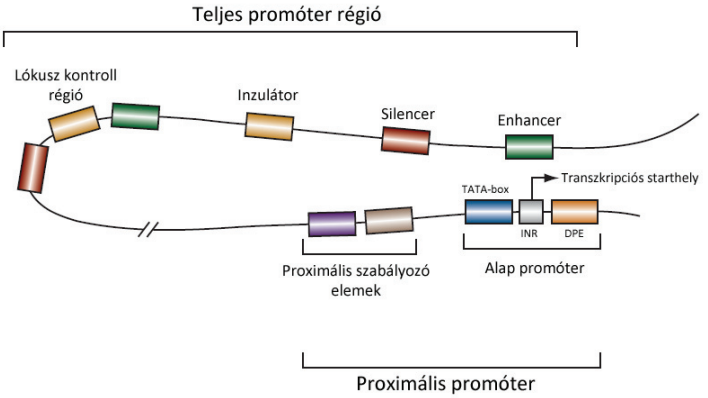
A különféle eukarióta szervezetek fiziológiás állapotának fenntartása, fejlődése, növekedése és túlélése igen sokrétű folyamat, több tízezer gén kifejeződését és folyamatos szabályozását igényli. A gének kifejeződésének irányítására nagyszámú különböző biológiai rendszer létezik, amelyek több szinten szabályozódnak. A szabályozás történhet a transzkripció folyamán, az mRNS módosításával, annak szállításakor, a transzláció alatt, vagy akár az mRNS stabilitásának változásával.

E folyamatok közül talán a legfontosabb a transzkripciós szabályozás. A transzkripció (átírás) során a gének RNS formába íródnak át a DNS-ről, és ez az RNS lesz egyrészt a transzláció (fehérjeszintézis) során az információszolgáltató, másrészt maga is különböző funkciókat tölthet be.

A transzkripció és a transzkripciós szabályozás pontos működésének felderítése régóta vizsgált genetikai, biokémiai és bioinformatikai kutatási téma. A transzkripció során a transzkripciós starthelynek (*transcription start site*, TSS) nevezzük a génnek a DNS-en található azon bázisát, amely elsőnek íródik át és az mRNS 5' végén található. A transzkripciós starthely körül található régiót alap promóternek (*core promoter*) hívjuk, mérete nem definiált pontosan, nagyjából 35 bázispár 5' vagy 3' irányban. Általában azt a transzkripciós starthely körüli régiót számítjuk ide, amely alapvetően szükséges a transzkripciós fehérjeapparátus adott génhez toborzásához, a különböző kiegészítő szabályozó faktorok mellett [1].

Az alap promóter mellett igen lényeges szerepet játszanak a transzkripciós szabályozásban a különféle egyéb szabályozó régiók és szekvencia motívumok is (1. ábra). Ezek közé tartoznak a proximális promóterek, különböző aktiváló és represszáló, vagy határoló/inzulátor elemek. Az aktiváló és represszáló elemek változatos helyeken találhatók, a géntől 5' vagy 3' irányban akár több 10 kilobázis távolságra, de egyes esetekben az 5' vagy 3' nem transzlálódó régiókban (*untranslated region*, UTR), esetleg magukban az exonokban vagy intronokban is előfordulnak. Ezek az elemek és régiók különböző szekvenciamotívumokat

tartalmaznak, amelyeket DNS kötő transzkripciós faktorok ismernek fel. A szekvenciamotívumok és főbb transzkripciós faktor családok részletesen ismertetésre kerülnek a következő fejezetekben.



1. ábra Az eukarióta promóter vázlatos struktúrája.

Az ábra vázlatosan mutatja az eukarióta promóter főbb régióit, az alap promótert, amelyben a transzkripciós starthely is megtalálható, 3 alap promóter motívumot, a proximális promótert, és több, a proximális promóter régión túl található szabályozó régiót. A különféle elemek részletes leírása a szövegben található. Az ábra a következő cikk 1. ábrája alapján készült: [2].

2.1.1. Az alap promóter és szekvenciamotívumai

Az alap promóterként meghatározott szekvenciarégióban több olyan motívum is megtalálható, amely központi fontosságú a minimális szintű transzkripció elindításában. Ezek a motívumok közvetlenül a transzkripciós fehérjeapparátus valamelyik tagjával lépnek kapcsolatba, amely az RNS polimeráz II mellett a transzkripciós faktor (TF) IIA, TFIIB, TFIID, TFIIE, TFIIF és TFIIH részekből áll. A TFIID fehérje több alegységet tartalmaz, a TATA-box kötő fehérje (*TATA-box binding protein*, TBP) mellett körülbelül 13 különböző transzkripciós faktor alkotja [3]. A felsorolt összetevők funkcionális egységgé összeállt formáját transzkripciós preiniciációs komplexnek (*transcription preinitiation complex*, PIC) is nevezzük. A

felsorolt transzkripciós faktorok közül elsősorban a TFIID és TFIIB rendelkezik lényeges szereppel a különböző motívumok felismerésében.

Ezek az alap promóterben megtalálható motívumok azonban nem általánosak és egységesen elterjedtek, mint az ezzel foglalkozó kutatások kezdeti időszakában gondolták, hanem nagyfokú strukturális és funkcionális variabilitást mutatnak [4] [5] [6]. Emellett valószínűleg ennek a variabilitásnak köszönhetően az alap promóter szerepet játszik a génexpresszió kombinatorikus szabályozásában és a transzkripciós válasz intenzitásának kialakításában is [7]. Alább következik a TATA-box, az INR elem és néhány további ismertebb és fontosabb motívum rövid ismertetése. A motívumok ismertetésénél azok konszenzus szekvenciáját az IUPAC kódtábla segítségével írtam le, amelynek részletes ismertetése a 2.4.1.1.-es fejezetben található.

Az alap promóter, és emellett a távolabbi promóter régiók is, bizonyos különbséget mutatnak a nagyobb élőlénycsoportok között. A gerincesek és zöld növények esetében több olyan strukturális jellemzője van a DNS-nek, amely lényegesen különbözik az elsődleges nukleotid szekvencia mellett [8]. Másrészt a különböző alap promóter motívumok eloszlása, pontos szekvenciája vagy megléte is eltérhet. Az eddigi kutatások nagy része azonban ember, egér, *Drosophila*, élesztő vagy hasonló, jól ismert modellfajokat vizsgált, és igen kevés adat áll rendelkezésre *Arabidopsis*, rizs vagy egyéb növényi promóter régiókból. Amennyiben ilyen eredmény megtalálható az irodalomban, azt külön jelölöm.

2.1.1.1. A TATA-box

A különböző alap promóter motívumok közül talán a TATA-box a legismertebb. A génszabályozással és promóter szerkezettel kapcsolatos egyik aránylag gyakori tévhit az, hogy minden promóternek TATA-box-ot kell tartalmaznia, amely azonban nem igaz, a TATA-box-ot tartalmazó génszabályozó régiók aránya a felfedezése óta folyamatosan csökken. Az eukarióta gének promóter régiójából elsőként írták le, mint DNS-kötő szabályozó elemet, Drosophila, emlős és vírus fehérjegének promóterének összehasonlító elemzése után [9] [10]. A TATA-box szinte az összes RNS polimeráz II által átírt gén előtt megtalálható volt, 25 – 30 nukleotid távolságra

5’ irányban a transzkripciós starthelytől. Élesztő génekben ez a távolság valamivel nagyobb, 40 – 100 nukleotid közé esik. A motívumot a már említett több egységből álló TFIID fehérjekomplex [11] TBP alegysége [12] ismeri fel, amelyet először élesztőben írtak le [13].

A központi konszenzus szekvencia kezdetben a TATAAA volt [14], azonban ez az újabb elemzések fényében módosult, és kiderült az is, hogy a TATA-box nem kötelező eleme az eukarióta promóternek [15]. Azon géneket vizsgálva, ahol a TATA-box egyáltalán előfordult, bioinformatikai módszerekkel sikerült a konszenzus szekvencia helyett egy pontosabb leírást lehetővé tevő pozíciónkénti nukleotid gyakoriságokat tartalmazó mátrixot generálni [16]. Pillanatnyilag talán a TATAWAAR konszenzus írja le az elemet a legpontosabban, ha a pozíciómátrixtól eltekintünk.

A TATA-box előfordulási arányaira adott becslések folyamatosan csökkentek, a Bucher és munkatársai [16] által végzett vizsgálatok alapján még a humán promóterek 78%-ában fordultak elő, amelynek oka valószínűleg a nem egyenletes mintavételezés volt. Későbbi vizsgálatok során, amikor az ENCODE projekt [17] keretein belül szintén humán promóterekben elemezték a funkcionális transzkripciós kötőhelyek gyakoriságát, ez 16 %-ra csökkent [18], de előfordulnak 5%-ot [8] vagy 2%-ot [19] mutató elemzések is.

A TATA-box szerepe és elhelyezkedése növények esetében is hasonló, mint az egyéb vizsgált fajoknál, gyakoriságuk *Arabidopsis* promóterek esetében 25% körül van[20], rizs promóterekben pedig 19% [21]. A rizs promóterekből leírt növényi TATA-box konszenzus szekvenciája CTATAWAWA az előző tanulmány alapján.

2.1.1.2. Az INR motívum

A TATA-box mellett egy másik, aránylag korán karakterizált alap promóter motívum az INR (*Initiator*) amely *cap signal* néven is ismert. A transzkripciós starthely környékén helyezkedik el, néhány nukleotidnyi kiterjedéssel 5’ és 3’ irányban. +1-es pozícióban a starthelynél adenozin, -1-es pozícióban általában citozin található, körülöttük pedig néhány pirimidin bázis (citozin, timin) [22]. Ezt az INR szekvenciát eltávolítva a transzkripció hatékonysága csökken, és a transzkripciós startpozíció elhelyezkedése sokkal heterogénebb lesz [23], amit először egy tengeri sün hiszton

H2A génnél figyeltek meg, egy az INR motívumot is érintő deléció után. Az INR és a TATA-box egymástól függetlenül is működőképesek, külön-külön is biztosítanak egy minimális szintű transzkripciót, azonban ha 25-30 nukleotid távolság van a két motívum között, hatásuk szinergisztikus [24] [25]. Élesztőnél a TATA-box-hoz hasonlóan, az INR motívum esetében sem szükséges a fent leírt aránylag pontos pozicionálás. Az INR motívumot szintén a TFIID komplex ismeri fel [26] [5], és ebben elsősorban a $TAF_{II}150$ és $TAF_{II}250$ transzkripciós faktorok játszanak szerepet, a TBP pedig a TATA-box – INR távolság pontos érzékelésénél döntő [27].

A központi aránylag jól konzervált CA nukleotidok [28] mellett a többi pozícióban kevéssé konzervált bázisok találhatók, kísérletes eredmények alapján az emlős konszenzus szekvencia az $YYA_{+1}NWYY$[1] [29], amivel a Drosophila szekvencia is egyezik [30]. A különféle promóter adatbázisok elemzése hasonló, bár nem teljesen egyező eredményt adott, emlősök esetében $YCA_{+1}NTYY$ [16], Drosophila esetében pedig $TCA_{+1}KTY$ [31]. Ezek mellett a jelentős variabilitást mutató motívumok mellett létezik egy úgynevezett „szigorú" (*strict*) INR motívum is, amely a $CCA_{+1}TYTT$ konszenzus szekvenciával írható le, és e központi motívum mellett további aránylag konzervált határoló régiókat is tartalmaz [32].

Valószínűleg az INR motívum fordul elő eukarióta fajoknál a leggyakrabban az alap promóter különböző motívumai közül, bár az INR esetében korántsem végeztek annyi átfogó elemzést, mint a TATA-box-nál. Egy ~2000 Drosophila alap promótert tartalmazó szekvencia kollekció vizsgálata során azok 62,8 %-ában találták meg az INR motívumot [31]. Két másik vizsgálat humán gének 85 %-ában [33] és 46 %-ában [34] találta meg az INR-t. Habár az eredmények itt is ellentmondásosak kissé, és valószínűleg befolyásolta őket a nem teljesen egyenletes mintavétel, olyan drasztikus mértékű csökkenés nem történt az előfordulások gyakoriságánál a vizsgálatok előrehaladtával, mint a TATA-box esetén.

Növények esetében az YYANWYY konszenzus csak limitált mennyiségben található meg, általánosságban azonban elmondható, hogy a transzkripciós starthely környékén erős preferencia figyelhető meg elsősorban a CA, másodsorban pedig a

[1] A konszenzus szekvenciában található +1 alsó index a transzkripciós starthelyet jelöli.

TA dinukleotid irányába [35]. Ez az YR konszenzussal leírható dinukleotid akár a 77%-ában is előfordul az *Arabidopsis* szabályozó régióknak.

2.1.1.3. A DPE motívum

A DPE (*Downstream Promoter Element*) gyakran fordul elő az INR mellett, elsősorban, de nem kizárólag TATA-box mentes promóterekben [36]. Kezdetben olyan elemként azonosították, ami TATA-box-ot nem tartalmazó promóterben a TFIID kötődéséhez szükséges. Drosophilától emberig szinte minden fajban konzerválódott, az INR motívumhoz képest pontosan meghatározott pozícióban, annak A_{+1} nukleotidjától +28-tól +32-ig lévő részen [37]. Az INR vagy a DPE motívum mutációja esetén a TFIID nem kötődik a promóterhez és a transzkripciós aktivitás megszűnik. A két elem közötti pontos távolság megváltozása esetén pedig drasztikusan csökken a transzkripció mértéke [36]. A DPE motívummal rendelkező promóterek az INR mellett általában nem tartalmaznak más alap promóter motívumot.

A DPE konszenzus szekvenciája talán az RGWYV szekvenciával írható le, amiből egyértelműen látszik, hogy az előző motívumokhoz képest sok nukleotid kombináció funkcióképes. Humán és egér promóterek elemzésével sikerült egy valamivel pontosabb pozíciónkénti nukleotid gyakoriságokat tartalmazó mátrixot előállítani [38].

A DPE gyakorisága Drosophila promóterekben 205 alap promóter vizsgálata alapján a következő: 26 % tartalmaz DPE elemet, 14 % pedig DPE-t és TATA-box-ot [37]. A DPE pozíciónkénti bázisgyakoriságait is megállapító, humán és egér szekvenciák jóval átfogóbb vizsgálata során a DPE mintegy 12 %-ában fordul elő a promóter régióknak [38].

2.1.1.4. A BRE motívum

A BRE (*TFIIB Recognition Element*) az egyetlen olyan alap promóter elem, amihez nem a TFIID komplex kötődik, hanem a TFIIB. A TFIIB – TBP – TATA-box komplex kristályszerkezete alapján már korábban nyilvánvaló volt, hogy a TFIIB és a TATA-box körüli DNS szekvencia között valamilyen interakció van [39]. Először olyan elemként definiálták, amely promóterek egy bizonyos részénél a TATA-box-tól 5’ irányban

helyezkedik el [40]. Későbbi kutatások során sikerült kimutatni egy, a TATA-box-tól 3' irányban elhelyezkedő BRE elemet is [41], és a BRE elnevezése BREu (*upstream*) és BREd (*downstream*) lett, azonban a szakirodalomban sokszor csak a BREu szerepel, egyszerűen mint BRE elem. A TATA-box-al együtt működnek, és növelhetik vagy csökkenthetik is a transzkripció mértékét.

A BREu konszenzus szekvenciája SSRCGCC, a BREd konszenzus szekvenciája pedig RTDKKKK a már említett elemzések alapján, azonban aránylag kevés tanulmány készült, ez a konszenzus szekvencia valószínűleg nem teljesen pontos, bár a BREu esetében létezik egy pozíciónkénti nukleotid gyakoriságot tartalmazó mátrix is [38].

A BREu elem gyakorisága humán és egér promóterek vizsgálata alapján 22% [38] körüli, egy másik elemzés pedig 25% [42] körüli értéket állapít meg. A BREd a BREu elemhez hasonló gyakorisággal fordul elő [41].

2.1.1.5. Egyéb alap promóter motívumok

Az MTE elemet (*Motif Ten Element*) Drosophila alap promóterek vizsgálata során azonosították, mint egy felülreprezentált funkcionális szekvencia motívumot [31]. Az INR elem A nukleotidjához képest +18 - +27 pozícióban helyezkedik el [43], a promóterek több mint 50%-ában megtalálható [38].

A DCE elemet (*Downstream Core Element*) először az emberi β-globin gén promóterében azonosították [44]. A transzkripciós starthelytől számított +10 - +45 pozíciók között helyezkedik el, és hozzájárul a TFIID kötődéséhez, valamint a transzkripciós aktivitáshoz.

A DRE (*DNA Replication-related Element*) olyan alap promóter motívum, amelyet nem a TFIID vagy TFIIB ismer fel, hanem egy a TBP-vel homológiát mutató fehérje, amelyet először Drosophila-ban azonosítottak és a TRF1 nevet kapta (*TBP Related Factor*). Rokon fehérjék a TRF2 és a TR3, az elsőt több különböző fajban megtalálták, a másodikat csak gerincesekben [45]. Ezek a TRF fehérjék kötődnek a DRE elemhez [46].

Az XCPE1 (*X gene Core Promoter Element 1*) elem aránylag frissen azonosították a hepatitis B vírus X génjének promóteréből [47]. Humán génekben a TATA-box-ot

nem tartalmazó gének egy részénél is megtalálható, *in vitro* elemzések szerint a TBP-t vagy a TFIID komplexet is hasznosítani tudja.

A PSE (*Proximal Sequence Element*) snRNS-ek alap promóterében található meg, meghatározza a transzkripciós starthelyet, és szükséges az alapvető transzkripcióhoz. Több eukarióta fajban leírták, RNS polimeráz II és III által átírt génekben is rendelkezik valamilyen szereppel [48].

A GA és CA elemek, amelyek az adott dinukleotidról lettek elnevezve, néhány éve leírt növényi alap promóter elemek, pontos funkciójuk és jellemzésük még kevéssé ismert [20]. A GA elem elsősorban Y-folt és INR elem mentes promóterekben fordul elő, a CA elem pontos eloszlása a kis mennyiségű rendelkezésre álló eredmény alapján még nem leírható. Más tanulmányok szerint a GA elem a TATA-box-al sem fordul elő közösen[49].

2.1.2. CpG szigetek

Az eukarióta promóterek egyik jellegzetessége a CpG szigetek jelenléte [50], amelyek olyan 0,5 – 2 kilobázis méretű DNS szakaszok, amelyek aránylag magas CpG dinukleotid tartalommal rendelkeznek. A genomban megtalálható CpG dinukleotidok jelentős része metilált a citozin 5. szénatomján [51], azonban a CpG szigetek dinukleotidjaira ez nem jellemző. Kezdetben a háztartási gének promóteréhez kapcsolták a CpG szigeteket, ez azonban a későbbi elemzések alapján megdőlni látszik, habár az irodalom nem egységes, sok esetben valószínűleg a nem megfelelő mintavétel, a pontatlan szekvencia adatok és inkomplett EST szekvenciák vezettek ellentmondásos eredményre.

A CpG szigetek általában nem tartalmaznak TATA-box-ot, INR vagy DPE elemet [52], ezzel szemben gyakran előfordul bennük a GC-box motívum, amely egy az Sp1 transzkripciós faktor által felismert szekvencia. Feltételezhetően több gyenge alap promótert tartalmaznak, a transzkripció több helyről is elindulhat, egy akár 100 nukleotid hosszúságban elnyúló régióban a CpG szigeten belül.

1031 humán promóter elemzése alapján azok mintegy 50%-a tartalmaz CpG szigetet [33], egy másik elemzés a transzkripciós starthely körüli ±100 nukleotid hosszú régiót vizsgálva a gének mintegy 70%-ánál talált CpG szigetet [53].

2.1.3. Növényi Y-foltok

A növényi promóterekben nem található meg a gerincesekre jellemző CpG szigetek, egy másik Y-folt névre hallgató elem viszont igen. A tipikus Y-folt C és T nukleotidokat (pirimidineket) tartalmaz, és csak növényekben fordul elő [54]. Előfordulása a transzkripciós starthelytől 13 nukleotidra 5' irányba mutat maximumot, azonban még 60 nukleotid távolságra is jelentős a mennyisége [35]. Biokémiai szerepe még nem ismert pontosan, rizs és *Arabidopsis* fajokban a promóterek mintegy 50%-ában fordul elő [20] [21].

2.1.4. Proximális promóter régió és szabályozó elemei

Az alap promóterről a témával foglalkozó kutatások kezdetén az volt a feltételezés, hogy csak egy általános, alap transzkripciós szintet határoz meg, és a különféle egyéb transzkripciós faktorok és azok kötőhelyei szabályozzák az expresszió pontos mértékét. Habár ma már úgy tűnik, hogy az alap promóter is igen sokrétűen tudja befolyásolni a transzkripciót, a különféle, tőle 5' vagy 3' irányban elhelyezkedő funkcionális elemeknek is nagy szerepük van [2].

Ezek a különböző aktiváló, represszáló és határoló szekvenciák, amelyekhez fehérjék kötődve az eukarióta szervezetek szöveteit alkotó sejtekben a génexpressziót modulálják és ezáltal szövetspecifikus, térben és időben változó mintázatot generálnak. A proximális promóter régiónak nevezett DNS szakasz, ahol ezek az elemek elhelyezkednek több 100 vagy 1000 nukleotid hosszúságú lehet, jellemző mérete nincs, maga a régió pedig akár több 10 kilobázis távolságra is elhelyezkedhet a promótertől.

A proximális promóter régiók esetében azonban jelentős és a későbbiekben lényeges tulajdonság az, hogy növények és állatok esetében a kódoló régióktól eltérően, a nem kódoló részek szubsztitúciós aránya különbözik. Növényeknél ez jóval magasabb, a promóterek szekvenciája, és ebből következően a transzkripciós faktorok kötőhelyei is degeneráltabbak, mint az állatok esetében [55].

2.1.4.1. Transzkripciós aktivátorok

Általánosságban elmondható hogy az aktivátor szekvenciákhoz kötődő fehérjék szekvencia-specifikus DNS kötő képességgel rendelkeznek, és felismerő helyeik megtalálhatók a promóter régiókban. Több különböző csoportjuk van, elsősorban a fehérje DNS kötő régiója alapján megkülönböztetve. A DNS kötő régió mellett általában fontos szerepet játszik egy úgynevezett aktivátor régió is, amely a transzkripciós folyamat stimulációjához szükséges. Amíg a DNS kötő régiók aránylag jól karakterizáltak a szakirodalomban funkcionálisan és strukturálisan is, az aktivátor régiókról ez nem mondható el.

Egy átfogó elemzés alkalmával, ami a humán transzkripciós faktorokat próbálta karakterizálni és katalogizálni, 1391 genomi lókuszról állapították meg nagy biztonsággal hogy transzkripciós faktor [56]. Ez a fehérjekódoló gének mintegy 6%-a. A valós érték valószínűleg 1700 és 1900 között található.

A DNS kötő régió alapján csoportosított aktivátorok az alábbi 8 nagy kategóriába sorolhatók [57] 240 fehérje térszerkezet elemzése alapján:

1. HTH (*helix-turn-helix*)
2. cinkujj (*zinc-coordinating*)
3. cipzár típusú (*zipper-type*)
4. egyéb α-hélix (*other α-helix*)
5. β-lemez (*β-sheet*)
6. β-hajtű/szalag (*β-hairpin/ribbon*)
7. egyéb (*other*)
8. enzim (*enzyme*)

A 8. csoport valójában kivétel a szerkezeti csoportosítás szempontjából, mivel minden olyan fehérjét tartalmaz amely enzimatikus aktivitást is mutat. E kategóriákon belül további csoportosítás lehetséges, amelynek eredménye 54 különböző fehérjecsalád, amelyből 33 egynél több fehérjét tartalmaz, ha az eredeti 240 fehérjéből álló gyűjteményt vesszük alapul.

A szerkezeti alapon történő csoportosítás igen hasznos annak megértésében, hogy egyes régiók miként kötődnek a specifikus DNS motívumokhoz, másrészt betekintést nyújtanak evolúciójukba is. Emellett bizonyos esetekben a DNS kötő régió alapján a transzkripciós faktor funkciójára is lehet következtetni, a HTH csoportba tartozó homeodomain régiót is hordozó transzkripciós faktorok általában egyedfejlődéssel kapcsolatos folyamatok szabályozásában vesznek részt, az interferon típusú transzkripciós faktorok pedig virális fertőzések elleni immunválasz beindításában játszanak szerepet [56].

Annak ellenére, hogy a transzkripciós faktorok DNS kötő régiói aránylag változatosak, alapvetően 3 féle régió dominál, a C_2H_2 cinkujj (cinkujj csoport), a homeodomain (HTH csoport) és a hélix-hurok-hélix (cipzár típusú csoport). Ez a 3 régió felelős a DNS kötésért a transzkripciós faktorok mintegy 80 %-ában.

2.1.4.2. A DNS kötő régió által felismert motívumok

Egyes alap promóterben található transzkripciós faktor kötőhelyek már ismertetésre kerültek, emellett általánosságban a következő mondható el róluk. A fehérje által felismert rész az esetek nagy részében egy rövid, 6-12 nukleotid hosszú DNS szakasz, amelyen belül meg lehet határozni egy 4-6 nukleotidos központi fontosságú, mag régiót. A TRANSFAC adatbázisban található kötőhelyek elemzése alapján a mag régió hossza vitatott, abban az esetben, ha kísérletes bizonyíték nem áll rendelkezésre, a fajok közötti konzerváltságot alapul véve ez a régió akár 10-12 nukleotid hosszú is lehet [58]. Sok esetben az aktivátorok homo- vagy heterodimert formálnak, és ennek következtében kötőhelyeik is két fél kötőhelyből állnak, amelyet változó mennyiségű tetszőleges nukleotid választhat el egymástól.

A transzkripciós faktor kötőhelyek sok esetben degenerált szekvenciák, azaz vannak olyan pozícióik, ahol több különböző nukleotidot is képes felismerni adott fehérje. Ilyen esetekben kérdéses, hogy a fehérje ténylegesen az elsődleges nukleotid szekvenciát ismeri-e fel, vagy pedig valamilyen egyéb tulajdonság, például a DNS térszerkezet a döntő, amely meglepően konzervált lehet aránylag különböző nukleotid sorrendek esetén is [59].

A transzkripciós faktor kötőhelyek méretüknél és degeneráltságuknál fogva igen sokszor előfordulhatnak egy genomban, ami elméletben mind potenciális kötőhely a transzkripciós faktor DNS kötő régiója számára. A valóságban azonban e kötőhelyek nagy része nem elérhető, a pontosan szabályozott kromatin struktúra nem teszi őket hozzáférhetővé [60]. Emellett a transzkripciós faktorok eloszlása sem egyenletes a sejtmagban, bizonyos területeken nagyobb koncentrációban fordulnak elő, ez is befolyással van arra, hogy adott régió kötőhelyeihez milyen faktorok kötődnek ténylegesen. Ezeknek a transzkripciós gyárnak nevezett régióknak a kialakítását szintén a kromatin struktúra segíti [61] [62].

A kötőhelyek degeneráltsága hatással van a transzkripciós faktorok egyéb tulajdonságaira is, adott degenerált kötőhely különféle variációit felismerő transzkripciós faktornak változik a kötéserőssége a DNS-hez, ezáltal például a korai fejlődés során fontos koncentráció gradiensek kialakításában segíthet [63]. A kötőhely változása ezek mellett konformációs változásokat is okozhat az transzkripciós faktor aktivátor régiójában, így megváltoztatva annak aktivitását [64].

2.1.4.3. Transzkripciós represszorok vagy csendesítők

A represszorok az aktivátorokhoz hasonló szekvencia specifikusan a DNS-hez kötődő transzkripciós faktorok, azonban negatív hatásuk van a transzkripcióra, csökkentik annak mértékét vagy gátolják a teljes folyamatot. Hatásukat több különböző mechanizmussal is elérhetik miután kötődtek a DNS-hez. Egyes esetekben gátolják az aktivátor kötődését a közelében lévő kötőhelyre [65], vagy közvetlenül oda kötődnek be, ahova az aktivátor is kötődne [66]. A represszor nem csak közvetlenül a kötőhelyre való bekötődést akadályozhatja meg, hanem a kromatin struktúrát is módosíthatja, és ezáltal akadályozza meg az aktivátorok vagy egyéb transzkripciós faktorok bekötődését [67]. Utolsó lehetőségként pedig előfordulhat, hogy magának a transzkripciós iniciációs komplexnek a létrejöttét akadályozza a represszor az alap promóter területén [68].

2.1.4.4. Inzulátor (határoló) szekvenciák

Az aktivátor és represszor motívumok mellett fontos szereppel bírnak a proximális promóter régióban az úgynevezett inzulátor vagy határoló régiók. Ezek a régiók

segítenek adott DNS régió elszigetelésében a szomszédos területek szabályozó régióitól és transzkripciós aktivitásától, így a genomot különálló részekre darabolják fel, leszűkítve a transzkripciós faktor kötőhelyek hatóterületét és elválasztva egymástól a különbözőképpen kifejeződő géneket, vagy éppen alternatív promótereket. Két fő hatással rendelkeznek, egyrészt gátolhatják az aktivátor fehérjék és a promóter szekvenciák kölcsönhatását, másrészt megakadályozhatják a heterokromatikus régiók továbbterjedését, fenntartva a kromatin határokat. Ez a két fő hatás egyes esetekben szét is választható egymástól, mint például a csirke β-globin gén esetében [69]. Egy a Drosophila genom határoló szekvenciáit feltérképező átfogó vizsgálat alapján pedig jól látható a két felsorolt hatás érvényesülése [70] a teljes genomban.

2.1.4.5. Enhancerek (távoli aktivátorok) és silencerek (távoli csendesítők)

A távoli aktivátor és csendesítő elemek szerkezetüket illetően alapvetően megegyeznek a proximális promóter szabályozó elemeivel, a lényegi különbség a transzkripciós starthelytől való távolságban van. Mivel pozíció és orientáció függetlenek, akár több 10 kilobázisra is elhelyezkedhetnek a starthelytől, adott esetben nem csak 5' hanem 3' irányban is [71]. Sok esetben modulárisan működnek, adott promóterre különböző időpontokban vagy szövetekben más-más enhancerek és/vagy silencerek hatnak [72].

Az elemek ilyen távolságból való működését a DNS hurok mechanizmus [73] teszi lehetővé. Ennek során a DNS flexibilitását kihasználva, a DNS-hez kötött távoli aktivátor vagy csendesítő elemek az alap promóter közvetlen közelébe kerülnek, és így kölcsönhatásba léphetnek a transzkripciós iniciációs komplex-el, egyéb transzkripciós faktorokkal vagy az alap és proximális promóter DNS-ével.

2.1.5. Alternatív promóterek és transzkripciós starthelyek

Habár a transzkripciós faktor kötőhelyek karakterizálásához szigorúan véve nem kapcsolódik, röviden szót kell ejteni az alternatív promóterekről és transzkripciós starthelyekről, az elmúlt néhány év egyik legfontosabb felfedezéséről a transzkripcióval kapcsolatban. Elsősorban humán és egér mintákat vizsgálva úgy tűnik, hogy a transzkripció nagyságrendekkel komplexebb folyamat és jóval több

transzkript képződik mint feltételeztük. Az átfedő vagy éppen antiszenz transzkriptek mellett [74] rengeteg az alternatív transzkripciós starthely, amelyek különböző promóter típusokat definiálnak eloszlásuktól függően [75]. Az alternatív promóterek száma is jelentős, egyes vizsgálatok szerint a humán gének 20-30 %-a rendelkezhet több elkülönülő szabályozó régióval [76].

2.2 Promóter adatbázisok

A szekvenálási módszerek fejlődésének köszönhetően a különféle publikus szekvencia adatbázisok ma már igen nagy mennyiségű szekvenciát tartalmaznak a legkülönfélébb fajokból. Az egyik legnagyobb szekvencia gyűjtemény, az Európai Nukleotid Archívum (*European Nucleotide Archive*, ENA) a legutolsó adatok szerint ~500 milliárd ($5*10^{11}$) nyers és illesztett szekvenciát tartalmaz, amelyek mintegy ~50 trillió ($5*10^{13}$) nukleotidot jelentenek [77]. Ezen szekvencia adatok feldolgozása azonban komoly számítási kapacitást és bioinformatikai háttértudást igényel. A különböző promóter szekvenciák, esetleg ortológ promóter csoportok kinyerése nem egyszerű annak ellenére, hogy egyre több olyan genomböngésző és adatbázis áll rendelkezésre, amelyek bizonyos mértékben megkönnyítik ezeket a feladatokat [78] [79] [80].

Több olyan eukarióta promótereket, esetleg ortológ/paralóg promóter gyűjteményeket tartalmazó adatbázis létezik, amelynek adatai korlátozott mértékben felhasználhatók, habár nem nyújtanak általános, a lehető legtöbb fajra kiterjedő szekvencia adatokat és/vagy elemzéseket, másrészt felhasználó-barátságuk is hagy kívánnivalót maga után. Alább következnek a fontosabb promóter adatbázisok rövid leírásai, amelyek a gerinces (*Chordata*) vagy zöld növény (*Viridiplantae*) csoportokhoz tartozó fajok szekvenciáit gyűjtik össze.

2.2.1 Általános eukarióta adatbázisok

2.2.1.1. EPD

Az EPD (*Eukaryotic Promoter Database*) adatbázis [81] (http://www.epd.isb-sib.ch) egy annotált, nem redundáns eukarióta POLII promóter gyűjtemény, kísérletesen meghatározott transzkripciós starthelyekkel. Egyedi kísérletes eredmények és

genomannotáló projektek eredményei segítik a promóterek helyzetének pontos meghatározását. Az adatbázis utolsó verziója 4809 kísérletesen igazolt promótert tartalmaz, a transzkripciós starthelyhez viszonyított -9999-es nukleotidtól +10000-ig. Ezek mellett 13046 rizs promóter is megtalálható benne, amelyek egy előzetes annotáción és automatikus minőségellenőrzésen már átestek, azonban az EPD másik szekciójától eltérően nem rendelkeznek minden esetben szilárd kísérletes bizonyítékkal.

2.2.2. Gerinces fajok promótereit tartalmazó adatbázisok

2.2.2.1. CORG

A CORG (*COmparative Regulatory Genomics*) promóter elemző keretrendszer [82] (http://corg.eb.tuebingen.mpg.de) 5 különböző faj szekvenciáit tartalmazza (ember, egér, patkány, fugu, zebrahal). A kezdeti összehasonlító elemzések után 16127 ortológ promóter csoportot sikerült meghatározni az ENSEMBL adatbázis 31-es verziója alapján. Az legújabb adatbázis már makákó, kutya, szarvasmarha, tyúk, gömbhal és karmosbéka szekvenciákat is tartalmaz, a fajok közötti konzervált régiók adataival kiegészítve.

2.2.2.2. DBTGR

A DBTGR (*DataBase of Tunicate Gene Regulation*) [83] (http://dbtgr.hgc.jp) különféle előgerinchúros fajok promóter szekvenciáira és transzkripciós szabályozásukra koncentrál, azonban pillanatnyilag még csak 184 promóter szekvenciát tartalmaz, a hozzájuk kapcsolódó transzkripciós faktor kötőhelyekkel együtt.

2.2.2.3. DBTSS

A DBTSS (*DataBase of Transcription Start Sites*) adatbázis [84] (http://dbtss.hgc.jp) az egyik legnagyobb promóter szekvenciákat is tartalmazó gyűjtemény, amely eukarióta mRNS-ek pontos transzkripciós starthelyeit tartalmazza. A két faj (ember és egér) 31 különböző sejttípusából ~330 millió transzkripciós starthely körüli szekvenciát sikerült meghatározni a TSS Seq elnevezésű módszerrel [85]. Ezen eredmények alapján pontosan feltérképezhetők a promóter régiók, sok esetben az egy génhez tartozó alternatív promóterek is. Összesen 17879, a RefSeq [86]

adatbázisban megtalálható gén transzkripciós starthelyeit és promótereit sikerült így leírni.

2.2.2.4. HemoPDB

A HemoPDB (*Hematopoiesis Promoter Database*) adatbázis [87] (http://bioinformatics.wistar.upenn.edu/HemoPDB) vérképzéssel kapcsolatos humán gének promótereit és az azokhoz kapcsolható transzkripciós faktor kötőhelyek adatait tartalmazza. Jelenleg 246 promóter szekvenciáról rendelkezik információval.

2.2.2.5. MPromDb

Az MPromDb (*Mammalian Promoter Database*) [88] (http://gdvtk.wistar.upenn.edu) egy minőségellenőrzésen átesett ChIP-Seq [89] kísérletes eredményeket tartalmazó gyűjtemény, amely bioinformatikai módszerekkel megjósolt és ismert aktív RNAP-II promótereket tartalmaz. Az adatbázis 26 különböző, az NCBI GEO [90] adatbázisból származó ChIP-Seq kísérletből ember esetében 6, egér esetében pedig 10 különböző sejt/szövettípus adatait integrálja. A végeredményként kapott 42893 humán és 48366 egér promóter további adatbázisokkal és szekvencia adatokkal lett integrálva.

2.2.2.6. OMGProm

Az OMGProm (*Orthologous Mammalian Gene Promoters*) adatbázis [91] (http://bioinformatics.wistar.upenn.edu/OMGProm) alapvetően humán és egér promóter régiók összehasonlító elemzésére szolgál. Több különböző adatbázisból származó eredményeket integrál, egyrészt kísérletesen vizsgált cDNS, promóter és első exon szekvenciákat, másrészt az NCBI HomoloGene homológia információit, harmadrészt pedig humán és egér genomi szekvenciákat. Jelenleg 8550 promóter párt tartalmaz 6373 humán és egér gén alapján.

2.2.2.7. PromoSer

A PromoSer [92] (http://biowulf.bu.edu/zlab/PromoSer) adatbázis lehetővé teszi humán, egér és patkány promóter szekvenciák automatizált letöltését és feldolgozását egy webes felületen keresztül. EST, mRNS és RefSeq adatokat integrál a transzkripciós starthelyek pontos meghatározásához és az EPD adatbázis eredményeit is tartalmazza.

2.2.2.8. TiProD

A TiProD (*Tissue-specific Promoter Database*) [93] (http://tiprod.bioinf.med.uni-goettingen.de) humán gének promótereit tartalmazza, a különbség a többi adatbázishoz képest abban van, hogy különféle expressziós adatok és annotációk alapján szövetspecifikus promóter gyűjtemények tölthetők le, összesen 52 kategóriában. A transzkripciós starthelyek meghatározása az EPD, DBTSS és ENSEMBL adatbázisok információin alapul.

2.2.3. Növényi fajok promótereit tartalmazó adatbázisok

2.2.3.1. Athena

Az Athena adatbázis [94] (http://www.bioinformatics2.wsu.edu/cgi-bin/Athena/cgi/home.pl) 30067 *Arabidopsis thaliana* promóter szekvenciát tartalmaz, amelyek különböző eszközökkel elemezhetők. A promóterek mellett 105 előzőleg már leírt transzkripciós faktor kötőhely konszenzus szekvenciája is megtalálható benne, amelyeknek vizsgálható a promóterekben való eloszlása és különböző tulajdonságai.

2.2.3.2. Osiris

Az Osiris [95] (http://www.bioinformatics2.wsu.edu/cgi-bin/Osiris/cgi/home.pl), hasonlóan épül fel, mint az Athena adatbázis, azonban a rizs genomot és a hozzá kapcsolódó expressziós adatokat, transzkripciós faktor kötőhelyek adatait dolgozza fel, 24209 gén szabályozó régióját tartalmazva.

2.2.3.3. PlantProm

A PlantProm [96] (http://mendel.cs.rhul.ac.uk/mendel.php?topic=plantprom) egy nem redundáns növényi promóter adatbázis, különböző fajokból, minden esetben kísérletes bizonyítékokkal a pontos transzkripciós starthelyre. 305 promóter régiót tartalmaz a -200-tól +50-ig terjedő starthely körüli régióból.

2.2.3.4. ppdb

A ppdb (*plant promoter database*) [97] (http://ppdb.agr.gifu-u.ac.jp/ppdb) az elérhető teljes genom szekvencia alapján *Arabidopsis thaliana* és rizs promótereket és azok annotációját tartalmazza, transzkripciós starthely információk mellett.

2.3. Transzkripciós faktor és transzkripciós faktor kötőhely adatbázisok

A különféle promóter adatbázisok mellett rengeteg olyan forrás is létezik, amelyek kifejezetten transzkripciós faktorok vagy transzkripciós faktor kötőhelyek információit tartalmazzák, megkönnyítve ezzel a transzkripciós szabályozás bioinformatikai vizsgálatát. Tartalmukat illetően igen eltérőek, egyes adatbázisok csak egy adott fajra koncentrálnak, mások pedig általános eukarióta kötőhely gyűjteményként funkcionálnak. Sok esetben a kísérletes adatokat bioinformatikai módszerekkel megjósolt kötőhelyek, funkcionális régiók is kiegészítik. A fontosabb adatbázisok és gyűjtemények felsorolása alább következik.

2.3.1. Általános eukarióta adatbázisok

2.3.1.1. JASPAR

A JASPAR [98] (http://jaspar.genereg.net) az aktuálisan elérhetőek közül az egyik legjobb minőségű, ellenőrzött és kevés redundáns adatot tartalmazó transzkripciós faktor kötőhely adatbázis. A legutolsó, negyedik verzió a 457 transzkripciós faktor kötőhely profil mellett egyéb adatbázisok és kísérletek eredményeit is feldolgozza. Gerinces és növényi szekciói mellett fonálféreg, rovar és gomba fajok adatai is megtalálhatók.

2.3.1.2. ooTFD

Az ooTFD (*object-oriented Transcription Factor Database*) [99] (http://www.ifti.org/ootfd) annak ellenére, hogy hosszabb ideje nincs fejlesztés alatt és nem frissülnek az adatai, hasznos információkat tartalmaz a transzkripciós faktorok mellett a több egységből felépülő transzkripciós faktor komplexeket illetően is.

2.3.1.3. ORegAnno

Az ORegAnno (*Open REGulatory ANNOtation database*) [100] (http://www.oreganno.org) egy nyitott, közösségi alapon és hozzájárulással működő, transzkripciós szabályozásra koncentráló adatbázis. Az első verzióban a már említett cisRED és az ENSEMBL adatbázis szolgáltatta az adatokat, ez az aktuális verzióban már jóval nagyobb és több mindent átfogó adattömegre terjed ki.

2.3.1.4. TRANSFAC és TRANSCOMPEL

A TRANSFAC és TRANSCOMPEL [101] (http://www.gene-regulation.com) az egyik legnagyobb transzkripciós faktor kötőhelyeket és transzkripciós faktorokat összegyűjtő adatbázis, rendszeres, évente többszöri frissítéssel. Az adatbázisnak egy aránylag kevés és nem túl friss adatokat tartalmazó ingyenes verziója mellett van egy éves előfizetési díjért elérhető jóval bővebb verziója, ami rengeteg információforrást és irodalmi adatot összegez, az egyik legátfogóbb ilyen típusú adatbázissá téve a TRANSFAC-ot.

2.3.1.5. TRRD

A TRRD (*Transcription Regulatory Regions Database*) [102] (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd) adatbázis a TRANSFAC-hoz hasonlóan rengeteg különböző irodalmi adatot integrál, génektől és kötőhelyektől kezdve különböző expressziós adatokig. Az összes fontosabb vizsgált eukarióta fajról található benne információ, az embertől és *Arabidopsis*-tól kezdve az ecetmuslicáig bezárólag.

2.3.2. Gerinces fajok kötőhelyeit tartalmazó adatbázisok

2.3.2.1. ABS

Az ABS adatbázis (*Annotated regulatory Binding Sites*) [103] (http://genome.crg.es/datasets/abs2005) 211 manuálisan feldolgozott és ellenőrzött promótert tartalmaz ember, egér, patkány és csirke fajokból, amelyek 100 ortológ csoportba sorolhatók. A promóterek transzkripciós starthelyeit a DBTSS segítségével sikerült meghatározni, a különböző kötőhelyek pedig a JASPAR [98] és TRANSFAC [101] adatbázisok adatai segítségével lettek megállapítva.

2.3.2.2. cisRED

A cisRED [104] (http://www.cisred.org) platform kihasználva a rendelkezésre álló nagymennyiségű genomi adatot, több különböző módszerrel és programmal elemzi az ortológ promóter csoportokat az ENSEMBL adatbázis teljes genomillesztései alapján. Humán, egér és patkány eredmények mellett aktuálisan már *C. elegans* adatokat is tartalmaz.

2.3.2.3. TRED

A TRED (*Transcriptional Regulatory Element Database*) [105] (http://rulai.cshl.edu/TRED) adatbázis az emlős szekvenciákra összpontosít, ezen belül is a humán, egér és patkány promóterekre. A promóterek két forrásból származnak, egyrészt az EPD és DBTSS adatbázisok adatait integrálja, másrészt a FirstEF promóter kereső program eredményeit is felhasználja, amelyet mRNS és EST szekvenciák is megerősítenek, az ortológ szekvenciák összehasonlítása mellett. 1900 humán promóter mellett mintegy 300 egér és patkány szekvenciát tartalmaz.

2.3.3. Növényi fajok kötőhelyeit tartalmazó adatbázisok

2.3.3.1. AGRIS

Az AGRIS (*Arabidopsis Gene Regulatory Information Server*) [106] (http://arabidopsis.med.ohio-state.edu) az *Arabidopsis* genom alapján 3 adatbázisba gyűjti a genomban található transzkripciós faktor géneket, a jósolt és kísérletesen igazolt transzkripciós faktor kötőhelyeket és a köztük lévő kölcsönhatásokat. 1773 transzkripciós faktor leírása mellett több mint 33000 darab 3000 nukleotid hosszúságú promóter és ~8100 gén – transzkripciós faktor kölcsönhatás kereshető és vizsgálható a webes felületen vagy tölthető le.

2.3.3.2. AthaMap

Az AthaMap [107] (http://www.athamap.de) lehetséges *Arabidopsis* transzkripciós faktorokat tartalmaz. 115 transzkripciós faktor leírása mellett a hozzájuk tartozó kötőhelyek pozíció specifikus mátrixai vagy konszenzus szekvenciái is megtalálhatók.

2.3.3.3. DATF

A DATF (*Database of Arabidopsis Transcription Factors*) [108] (http://datf.cbi.pku.edu.cn) adatbázis a harmadik, amely *Arabidopsis* transzkripciós faktorokkal foglalkozik, a különféle transzkripciós faktor családokról, azok struktúrájáról és filogenetikai kapcsolatairól is részletes elemzést ad, az irodalmi adatok és annotációk összegyűjtése mellett. 2290 gén modellt tartalmaz 64 családba sorolva.

2.3.3.4. PLACE

A PLACE (*Plant cis-acting regulatory DNA elements*) [109] (http://www.dna.affrc.go.jp/htdocs/PLACE) az egyik legrégebbi növényi transzkripciós faktor kötőhelyeket tartalmazó adatbázis. Már nem frissítik, azonban sok különböző adatbázis és program felhasználja az eredményeit, az utolsó verzió 208 különböző kötőhelyet tartalmaz.

2.3.3.5. PlantCARE

A PlantCARE [110] (http://bioinformatics.psb.ugent.be/webtools/plantcare/html) 435 transzkripciós faktor kötőhely információjával a PLACE mellett a másik széles körben ismert növényi transzkripciós faktor kötőhely adatbázis, szolgáltatásait a mai napig használják a már nem túl felhasználóbarátnak számító felület és a frissítések hiánya ellenére.

2.3.3.6. PlantTFDB

A DATF adatbázis és több hasonló, de más fajokra koncentráló webes szolgáltatás összevonásával és integrálásával jött létre a PlantTFDB (*Plant Transcription Factor Database*) [111] (http://planttfdb.cbi.pku.edu.cn) adatbázis, amely 49 növényfaj transzkripciós szabályozással kapcsolatos adatait tartalmazza.

2.3.3.7. PlnTFDB

A PlnTFDB (*Plant Transcription Factor Database*) [112] (http://plntfdb.bio.uni-potsdam.de/v3.0) a PlantTFDB-hez igen hasonló adatbázis minden olyan tágabb értelemben vett növényfaj transzkripciós szabályozáshoz kapcsolódó információit és szekvencia adatait próbálja feldolgozni, amely teljesen megszekvenált genommal rendelkezik. Az egysejtű vörös és zöld algáktól a virágos növényekig bezárólag 19 fajt, és 84 transzkripciós faktor családot tartalmaz.

2.4. 'In silico' transzkripciós faktor kötőhely predikciós módszerek

A transzkripciós faktor kötőhelyek keresése és előrejelzése az elmúlt években az egyik legnagyobb kihívást jelentő bioinformatikai feladat lett. A lassan átláthatatlan mennyiségű különböző program sokszor igen eltérő megközelítéssel, statisztikai módszerekkel és algoritmusokkal próbálja a hasonlóan szabályozott gének

promótereinek feltételezett kötőhelyeit megtalálni. Az alábbiakban a jellegzetesebb módszereket és programokat próbálom áttekinteni a teljesség igénye nélkül.

A különféle kötőhelyek előrejelzése alapvetően két nagy csoportra osztható. Egyrészt kiindulhatunk kísérletes adatokból, amik alapján hasonló, feltételezhetően szintén biológiai szereppel rendelkező szekvencia motívumokat keresünk egy adott promóterben, másrészt teljesen új motívumokat is kereshetünk expressziós profiljuk alapján feltételezhetően hasonlóan szabályozott, esetleg ortológ gének szekvenciáiban.

2.4.1. Ismert kötőhelyek keresése

Az ismert transzkripciós faktor kötőhelyek keresése a kevésbé komplikált problémák közé tartozik, a programok nagy része egyszerű konszenzus szekvenciát vagy esetleg pozíció specifikus súlymátrixot alkalmaz.

2.4.1.1. Konszenzus szekvenciák

A különféle kötőhelyek reprezentálására a különféle konszenzus szekvenciák a legegyszerűbb módszer, és ezek okozzák a legkevesebb problémát is a keresőprogramok számára. A konszenzus leírására a standard IUPAC (*International Union of Pure and Applied Chemistry*) kód szolgál, amelynek részleteit az 1. táblázat tartalmazza.

A különféle konszenzus szekvenciák sok területen felhasználhatók, különböző szekvencia illesztések, mRNS másodlagos struktúra előrejelzések, fehérjék távoli rokonságának megállapítása során, és természetesen a transzkripciós faktor kötőhelyek jellemzésére is. Több módszer és szabály létezik, ami alapján konszenzus szekvencia generálható [113], és a kapott eredménytől függően változik a keresés pontossága, szenzitivitása, amit elsősorban az engedélyezett nukleotid eltérések (*mismatch*) és a konszenzus szekvenciában található lötyögős (*ambiguous*) nukleotidok befolyásolnak [114].

2.4.1.2. Reguláris kifejezések

A konszenzus szekvenciákkal való keresés az elemezni kívánt adatmennyiségtől függően akár igen egyszerűen is megvalósítható. A Perl és Bioperl [115]

programozási nyelv és modulcsomag segítségével rövid programok írhatók amelyek reguláris kifejezések segítségével keresik meg az egyezéseket, azonban ez csak aránylag kis mennyiségű adat esetén hatékony.

Szimbólum	Jelentés	Komplementer	Magyarázat
A	A	T vagy U	**A**denin
C	C	G	**C**itozin
G	G	C	**G**uanin
T vagy U	T	A	**T**imin vagy **U**racil
M	A vagy C	K	a**M**ino
R	A vagy G	Y	pu**R**in
W	A vagy T	W	**W**eak (gyenge; 2 H kötés)
S	C vagy G	S	**S**trong (erős; 3 H kötés)
Y	C vagy T	R	p**Y**rimidine (pirimidin)
K	G vagy T	M	**K**eto
V	A, C vagy G	B	nem T vagy U
H	A, C vagy T	D	nem G
D	A, G vagy T	H	nem C
B	C, G vagy T	V	nem A
N vagy X	A, C, G vagy T	X vagy N	a**N**y (bármely)

1. táblázat IUPAC szimbólumok a nukleotidok szabványos jelölésére.

2.4.1.3. FUZZNUC program

Az EMBOSS [116] programcsomag FUZZNUC programja attól függően, hogy milyen típusú konszenzus szekvenciát kap, abban mennyi a feltételesen ismétlődő rész vagy

a bizonytalan bázis, más-más algoritmusokat használ a kereséshez, és nagyságrendekkel gyorsabb lehet, mint az egyszerű Perl scriptek segítségével végzett elemzés. A fent említett 2 aránylag sűrűn használt módszer mellett igen sok egyszerű program létezik, és ma már szinte minden bioinformatikai szoftvercsomagban implementálták valamilyen formában a konszenzus szekvenciákkal való keresést. Az egyszerű és hatékony módszer nagy hibája azonban, hogy rengeteg fals pozitív találatot adhat, és valójában információt veszítünk, amikor adott transzkripciós kötőhelyről rendelkezésre álló adatokat egy konszenzussá egyszerűsítjük.

2.4.1.4. Pozícióspecifikus mátrixok

A kötőhelyek pontosabb leírása konszenzus szekvenciák helyett különféle mátrixokkal is lehetséges, amelyek a kötőhely minden egyes pozíciójára tartalmazzák a lehetséges nukleotidok gyakoriságát vagy valamilyen származtatott (információtartalom, súlyozás) értéket. Mivel az irodalomban sok helyen felváltva használják a nem pontosan meghatározott elnevezéseket, az alábbiakban mindegyik mátrix típusról következik egy rövid leírás.

A gyakorisági mátrix (*position frequency matrix*, PFM)a legegyszerűbb, adott oszlopában tartalmazza a kötőhely egy pozíciójában a kísérletes eredmények alapján előforduló nukleotidok mennyiségét, esetleg egyszerű százalékszámítás alapján egy százalékos értéket vagy egy 0 és 1 közé eső gyakoriságot. Egy egyszerű gyakorisági mátrixot mutat az 1. egyenlet.

$$1.\ \begin{bmatrix} 4 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 3 & 3 & 0 & 2 & 1 \\ 0 & 0 & 1 & 2 & 1 & 1 \end{bmatrix} \begin{matrix} A \\ C \\ G \\ T \end{matrix}$$

A súlymátrix (*position weight matrix*, PWM) számított értékeket tartalmaz a gyakorisági mátrix alapján, és feltételezzük, hogy a mátrix pozíciói függetlenek egymástól. A súlymátrix számításának egyik módját a 2. egyenlet mutatja [117] az eredményét a képlet alapján pedig a 3. egyenlet. N a szekvenciák száma, p_i az adott bázis *a priori* valószínűsége, és $f_{i,j} = n_{i,j}/N$ az i bázis gyakorisága j pozícióban.

$$2.\ \ln\frac{(n_{i,j}+p_i)/(N+1)}{p_i} \approx ln\frac{f_{i,j}}{p_i}$$

$$3.\begin{bmatrix} 1.2 & 0 & -1.6 & 0 & -1.6 & 0 \\ -1.6 & -1.6 & -1.6 & 0 & 0 & 0 \\ -1.6 & 0.96 & 0.96 & -1.6 & 0.59 & 0 \\ -1.6 & -1.6 & 0 & 0.59 & 0 & 0 \end{bmatrix}\begin{matrix} A \\ C \\ G \\ T \end{matrix}$$

A dinukleotid súlymátrix (*dinucleotide weight matrix*, DWM) [118] az egyszerű súlymátrix egy továbbfejlesztett változata, ahol a nukleotid gyakoriságok helyett dinukleotid gyakoriságok írják le a kötőhelyeket, bizonyos mértékig így figyelembe véve azt, hogy a pozíciók nem függetlenek egymástól.

A pontozómátrix (*position scoring matrix*, PSM) kísérletes eredmények alapján egy empirikus értéket ad minden pozícióban a különböző bázisoknak [117].

A felsorolt mátrixokkal való keresésre több program alkalmas, a Perl programozási nyelv TFBS [119] modulcsomagja kezeli az összes típust. A TRANSFAC adatbázis, amely egy lokális webszerverre telepíthető programcsomagot is tartalmaz [101], szintén rendelkezik olyan programokkal, amelyek lehetővé teszik e mátrixokkal való keresést.

A MotifScanner program komplexebb módszert használ [120], a program alapját adó Gibbs mintavételezési eljárást [121] kombinálja egy komplex, magasabb rendű nukleotid háttérmodellel.

2.4.2. Ismeretlen kötőhelyek keresése

A DNS szekvenciákról felhasznált információ alapján az új, eddig még ismeretlen motívumokat kereső programokat három nagy csoportra oszthatjuk: 1) olyan módszerek, amelyek egy genomból a hasonlóan szabályozott gének promótereit használják fel, 2) módszerek, amelyek adott gén különböző fajokból származó ortológ promóter szekvenciái alapján keresnek motívumokat és 3) azok a módszerek, amelyek egy fajból származó hasonlóan szabályozott gének információi mellett ortológ gének filogenetikai információit is felhasználják [122].

Egy másik megközelítés szerint két nagy kategóriára lehet osztani a különböző motívumkereső programokat: 1) szó alapú (szabadszavas) keresést megvalósító

programok, amelyek az oligonukleotid gyakoriságok pontos meghatározására és összehasonlítására épülnek és 2) valószínűségi modellek, ahol a modell paramétereit maximum-valószínűségi vagy Bayes módszerekkel becsüljük.

A szó alapú keresőmódszerek igen gyorsak és hatékonyak lehetnek, különösen olyan esetekben, amikor egy pontosan meghatározott motívumot keresünk, amelynek a pozícióiban nincsenek bizonytalan bázisok. A valóságban azonban sokszor előfordulnak bizonytalan, gyengén konzervált helyek. Ennek köszönhetően általában szükség van valamilyen utólagos csoportosításra és szűrésre. A módszer másik nagy hibája a rengeteg hamis pozitív eredmény, ami biológiai szereppel nem rendelkezik.

A valószínűségi módszerek esetében a kötőhelyet egy pozíció specifikus súlymátrix írja le, ahol a motívum minden egyes pozíciójában ismert a nukleotidok kísérletekben megfigyelt gyakorisága. A mátrix megjelenítésére általánosan használt a szekvencia logó [123], amely adott pozíció nukleotid gyakorisága és konzerváltsága alapján arányos méretben ábrázolja az A, C, G, T nukleotidokat. Ezekkel a módszerekkel aránylag hatékonyan és kevés paraméter beállításával lehet keresni, de nagy hibájuk, hogy a különféle szabályozó régiók valószínűségi modelljein alapulnak. Ennek következtében aránylag kis változás vagy hiba a bemeneti adatokban jelentősen befolyásolhatja a végeredményt. Másrészt az sem biztosított, hogy a globálisan optimális megoldást adják, mivel sokszor valamilyen lokális keresést alkalmaznak, amelyek egy lokális optimumhoz fognak közelíteni.

A különféle motívumkereső módszerek tárgyilagos összehasonlítása és teljesítményük értékelése igen nehéz két kérdés miatt is. Egyrészt kevéssé ismert a szabályozó mechanizmusok és transzkripciós faktor kötőhelyek működésének pontos biológiai háttere, így nehéz azokat a kritériumokat megállapítani és modellezni, amelyek alapján eldönthető egy találat pontossága [124]. Míg a fehérjemodellezés esetén rendelkezésre állnak kristálystruktúrák, amelyek referenciaként szolgálnak, a kötőhelyeket előrejelző programok esetében semmi ilyen nincs pillanatnyilag. Másrészt a különféle módszerek különféle adatgyűjtemények esetében működnek optimálisan, a háttérben lévő igen változó

és összetett motívummodellek miatt, és emiatt sem lehetséges egy egységes referencia adatkészlet összeállítása.

2.4.2.1. Hasonlóan szabályozott gének promóterein alapuló módszerek

A kezdeti algoritmusok nagy része hasonlóan szabályozott gének promóter régióiban keresett statisztikailag felülreprezentált motívumokat. Az alapján, hogy pontosan milyen módszerre épülő algoritmust használ az adott program, további csoportosítás lehetséges.

2.4.2.1.1. Szó alapú algoritmusok

Az egyik legelső program, az Oligo-Analysis [125] segítségével hasonlóan szabályozott élesztő gének promóter régióit elemezték alkotói. A program a bizonyos mértékű egyszerűsítéseket használó heurisztikus algoritmusokhoz képest pontos és részletes elemzést végez. Ennek ára azonban az, hogy csak aránylag egyszerű és rövid szekvencia motívumokat képes megtalálni, másrészt minden pozíciónak egyeznie kell, nem lehetséges változó bázisú részek definiálása.

Az YMF (*Yeast Motif Finder*) [126] szintén részletes keresést végez, azonban megadható neki, hogy a motívum közepén maximum hány tetszőleges nukleotid legyen, így a dimerként kötődő transzkripciós faktorok kettős kötőhelyei is megtalálhatók. Emellett meghatározható az eltérő nukleotidok maximális száma is. Mindezek mellett a program felhasznál egy trinukleotid háttérmodellt is, amely a vizsgált faj promóter régiói alapján készül.

2.4.2.1.2. Valószínűségi algoritmusok

A valószínűségi algoritmusokat felhasználó módszerek különböző statisztikai módszereket használnak fel, mint a várható érték maximalizálás (*expectation maximalization*, EM), Gibbs mintavételezés (*Gibbs sampling*) vagy ezek valamely továbbfejlesztett verzióját.

A MEME [127] program a várható érték maximalizálás egy kiterjesztett verzióját használja fel, és segítségével nem illesztett biopolimer szekvenciákban keres egy vagy több ismétlődő motívumot.

A NestedMICA [128] a MEME algoritmusához hasonló, a várható érték maximalizálást használja, azonban a továbbfejlesztett mintavételezési eljárásnak, és egy új, mozaikos háttérmodellnek köszönhetően igen magas az érzékenysége.

A Gibbs mintavételezési eljárást használja fel az AlignACE [129], amely a vizsgált szekvencia gyűjteményben felülreprezentált motívum súlymátrixokat ad vissza eredményként. Figyelembe veszi a genom nukleotid gyakoriságát, a DNS mindkét szálán keres, nem engedélyezi az átfedő motívumokat és szekvenciánként csak 1 motívumot ad eredményként.

A MotifSampler [130] szintén egy módosított Gibbs mintavételezési eljárást alkalmaz, a két lényeges módosítás a következő: egyrészt a program egy valószínűségi eloszlást használ a szekvenciákban előforduló lehetséges motívumok számának megbecslésére, másrészt egy magasabb rendű Markov láncot is magába foglal, a háttér modellezésére.

2.4.2.2. Filogenetikai információt felhasználó módszerek

A filogenetikai információt felhasználó módszerek esetén az egyik alapvető feladat a megfelelő ortológ és paralóg viszonyok meghatározása. Ortológnak akkor tekinthetők gének, ha két vagy több különböző fajban találhatók, egy közös gén őstől származnak és ugyanazt a funkciót szolgálják. Paralógok abban az esetben, ha ugyanabban a fajban találhatók, és a közös gén ősből duplikáció útján jöttek létre. Ebben az esetben funkciójuk eltérhet egymástól, azonban általában összefüggésben vannak. A filogenetikai lábnyom módszer ortológ gének vizsgálatára alapul, azonban az evolúció során történt gén, kromoszóma, részleges vagy teljes genom duplikációk eredményeként az ortológ-paralóg viszonyok sokszor nem állapíthatók meg egyértelműen, ez a megfelelő szekvenciák összeválogatásában problémákat okozhat.

A módszerek közül a legnyilvánvalóbb valamilyen többszörös szekvencia illesztést végző program felhasználása. Az egyik legáltalánosabban elterjedt program a CLUSTALW [131], azonban globális szekvencia illesztő algoritmusa miatt promóterek konzervált régióinak keresésére nem a legideálisabb. Emellett az illesztés eredményei erősen függenek a szekvenciákat adó fajok evolúciós távolságától. Túl

közeli rokonok esetén az illesztés pontos, de nem informatív, túl távoli rokonoknál pedig igen nehéz a megfelelő szekvencia illesztés megtalálása.

A CONREAL [132] program arra a feltételezésre épül, hogy a funkcionális transzkripciós faktor kötőhelyek szekvenciája és sorrendje konzervált az ortológ promóter régiókban. Ez alapján a valamely ismert kötőhelyeket tartalmazó adatbázis információit felhasználva, a szekvenciákban lévő kötőhelyeket horgonyként használva végzi el a promóter régiók illesztését.

A PhyloCon [133] az ortológ szekvenciák mellett a hasonlóan szabályozott gének promótereit is figyelembe veszi. Első lépésként adott gén ortológjait illeszti és meghatározza a konzervált régiókat, ezzel génprofilokat képezve, majd a különböző gének profiljait hasonlítja össze. Ennek a végeredményei a feltételezett szabályozó motívumok.

3. Célkitűzések

Munkánk alapvető célja az volt, hogy a transzkripciós szabályozás bioinformatikai vizsgálatát megkönnyítsük és ehhez különböző eszközöket hozzunk létre, többek között egy minél több ortológ promóter csoportot tartalmazó adatbázist, amihez egy weben elérhető keresőfelületet is biztosítunk.

Az első feladat egy olyan módszer kifejlesztése volt, amellyel az egyes gének annotált első exonjainak a szekvenciáját felhasználva megkeressük a feltételezhetően ortológ első exonokat az aktuálisan elérhető genomi szekvenciákban. Ezután az ortológ első exonok előtti 500, 1000 és 3000 nukleotid hosszú DNS-szakaszokat tekintjük promóter régiónak, és helyezzük el ortológ géncsoportonként a DoOP adatbázisba (Database of Orthologous Promoters, http://doop.abc.hu). A humán és az *Arabidopsis* teljes genom annotáció alapján célul tűztük ki egy gerinces (*Chordata*) és zöld növényi (*Viridiplantae*) promótereket tartalmazó gyűjteményt létrehozását is.

Cél volt emellett a promóter gyűjtemények további vizsgálata és a feltehetően biológiai szereppel rendelkező evolúciósan konzervált régiók meghatározása a promóterekben. Különböző eszközöket fejlesztettünk ehhez a feladathoz is, és emellett vizsgáltuk a motívumok meghatározására leginkább alkalmas szekvencia illesztő és értékelő módszereket.

4. Anyagok és módszerek

4.1. Felhasznált számítógépek

A szekvencia összehasonlítások, BLAST keresések és elemzések, ahol nincs külön jelezve, egy Sun Fire V480R szerveren futottak. A szerver ~1 TB tárhellyel, 16 GB memóriával és 4 darab 900 MHz órajelű UltraSparc-III+ processzorral rendelkezett, az operációs rendszer pedig 64 bites Solaris 9 (SunOS 5.9) volt.

A kész adatbázisok, a hozzájuk fejlesztett webes keresőfelület és API egy dedikált Dell szerveren található, amely ~100 GB tárhellyel, 4 GB memóriával és 4 darab 2.4 GHz órajelen futó Intel Xeon processzorral rendelkezik, az operációs rendszer pedig OpenSUSE 11.1.

Egyéb esetekben az elemzések és a különféle szekvencia manipulációk egy 1,8 GHz órajelű Pentium 4 processzort tartalmazó, 1 GB memóriával és 500 GB tárhellyel rendelkező asztali számítógépen futottak, amelyen az operációs rendszer Gentoo Linux volt.

4.2. Programok és programozási nyelvek

Az adatbázis készítése és elemzése során rengeteg olyan kisebb-nagyobb feladat merült fel, amelyhez nem létezik szoftver, esetleg elavult, vagy nem megfelelően testre szabható az elérhető megoldás. Ilyen esetekben gyorsan megírható rövid programokat alkalmaztunk, amelyekhez a Perl (http://www.perl.org) programozási nyelv 5.8-as, később pedig 5.10-es verzióját használtuk fel.

Egyes feladatoknál, amelyek bioinformatikai elemzéseknél gyakran előfordulnak, már kialakult modulcsomagok léteznek azok rutinszerű megoldására. A Bioperl [115] modulcsomag alapvető segítséget nyújt a különféle biológiai adatformátumok kezelésében, konvertálásában és manipulálásában, munkánk során az egyik alapvető eszköz volt. A Bioperl mellett a konzervált régiók adatainak manipulálására a TFBS [119] modulcsomagot is felhasználtuk.

Több alkalommal, amikor a Perl és a Bioperl modulcsomag nem volt megfelelő teljesítményű, az EMBOSS [116] programcsomagot használtuk fel a szekvenciák manipulálására és szűrésére.

Az eredmények kiértékelésénél és ábrázolásánál az R statisztikai programozási nyelv [134] 2.10.0-s verzióját használtuk.

A keresőfelület létrehozásához és a weboldalak dinamikus generálásához a PHP (http://www.php.net) programozási nyelv 5.2-es verzióját használtuk, kombinálva a Perl CGI moduljával és a később leírásra kerülő, szintén Perl alapú DoOP API-val. Az adatok tárolására és indexelésére a MySQL adatbázis kezelő szoftver 5.0-s verzióját használtuk. A webes felület az XHTML 1.0 Strict és CSS 2 szabvány szerint készült, amelyhez a YUI (*Yahoo! User Interface*, http://developer.yahoo.com/yui) 2.6-os verziójú JavaScript és CSS könyvtára nyújtott segítséget.

4.3. Szekvencia adatbázisok

A munka folyamán több különböző publikus adatbázis szekvenciáit és genom annotációit is felhasználtuk.

4.3.1. Gerinces adatbázis szekvenciái

A gerinces adatbázis utolsó, aktuálisan elérhető verziójának (1.4) referencia szekvenciáit a humán genom adta, a későbbiekben tárgyalt ortológ promóter kereső módszernél az NCBI 36-os genom annotációjának 1-es verzióját használtuk 2006 februárból.

Az NCBI BLAST adatbázisai közül a gss (*genomic survey sequences*), htgs (*high throughput genomic sequences*), nt (*nucleotide collection*), wgs (*whole-genome shotgun reads*) és *other genomic* szekciók szekvenciáit használtuk fel, eltávolítva az összes mRNS (cDNS) szekvenciát és minden nem-gerinces szekvenciát is.

Az elérhető teljes genomszekvenciák közül az ENSEMBL adatbázisból töltöttük le kétféle zsákállat (*Ciona savignyi* és *Ciona intestinalis*), a zebradánió (*Danio rerio*), fugu (*Takifugu rubripes*), zöld gömbhal (*Tetraodon nigroviridis*), sarkantyús karmosbéka (*Xenopus tropicalis*), tyúk (*Gallus gallus*), törpeoposszum (*Monodelphis domestica*), patkány (*Rattus norvegicus*), egér (*Mus musculus*), szarvasmarha (*Bos*

taurus), kutya (*Canis familiaris*), rézuszmajom (*Macaca mulatta*) és csimpánz (*Pan troglodytes*) adatait.

A JGI (*DOE Joint Genome Institute*) oldaláról elérhető volt az egyik zsákállatfaj (*Ciona intestinalis*) genomszekvenciájának egy másik verziója.

A Broad intézet (*Broad Institute*) eredményei közül a következő genomokat használtuk fel: tüskés pikó (*Gasterosteus aculeatus*), kis sün tanrek (*Echinops telfairi*), kilencöves tatu (*Dasypus novemcinctus*), törpeoposszum (*Monodelphis domestica*), északi mókuscickány (*Tupaia belangeri*), erdei cickány (*Sorex araneus*), tizenhárom sávos ürge (*Spermohilus tridecemlineatus*), nyúl (*Oryctolagus cuniculus*), tengerimalac (*Cavia porcellus*), barna denevér (*Myotis lucifugus*), európai sün (*Erinaceus europaeus*), kutya (*Canis familiaris*), macska (*Felis catus*), elefánt (*Loxodonta africana*), északi óriás fülesmaki (*Otolemur garnettii*).

Az adatbázis elkészítése folyamán, ahol lehetséges volt, megállapítottuk a transzkripciós starthelyek pontos pozícióját egyéb független adatbázisok szekvencia adatai alapján. A DBTSS (*Database of Transcriptional Start Sites*) [84] 5.2-es verziójának humán, egér és zebradánió szekvenciáit használtuk fel a promóterek transzkripciós starthelyének meghatározására. Az EPD (*Eukaryotic Promoter Database*) [81] 88-as verziójából az összes gerinces adatot felhasználtuk, a PromoSer [92] adatbázis humán, egér és patkány szekvenciái mellett.

4.3.2. Növényi adatbázis szekvenciái

A növényi adatbázis esetén, mind a 3 elérhető verziónál (1.5, 1.6, 1.8) a referencia szekvenciák az *Arabidopsis* genom annotáció alapján lettek összegyűjtve. Az utolsó, 1.8-as verziójú adatbázis esetében a felhasznált genom verzió a TAIR8-as, 2008 júliusból.

Az NCBI BLAST adatbázisai közül a gerinces adatbázishoz hasonlóan a gss (*genomic survey sequences*), htgs (*high throughput genomic sequences*), nt (*nucleotide collection*), wgs (*whole-genome shotgun reads*) és *other genomic* szekciók szekvenciáit használtuk fel, eltávolítva az összes mRNS (cDNS) szekvenciát, és értelemszerűen kizárólag a zöld növényi szekvenciákat megtartva.

Az adatbázis elkészítésének időpontjában nem volt olyan növényi genom, amelynek szekvenciáit ne tartalmazta volna az NCBI BLAST adatbázisainak valamely felhasznált szekciója, így további genomok külön nem kerültek letöltésre.

A transzkripciós starthelyek pozíciójának megállapításához az EPD 99-es verziójának zöld növényi szekvenciáit és a PlantProm [96] adatbázis 2002.01-es verziójának adatait használtuk fel.

4.4. Humán és *Arabidopsis* keresőszekvenciák és exon típusok

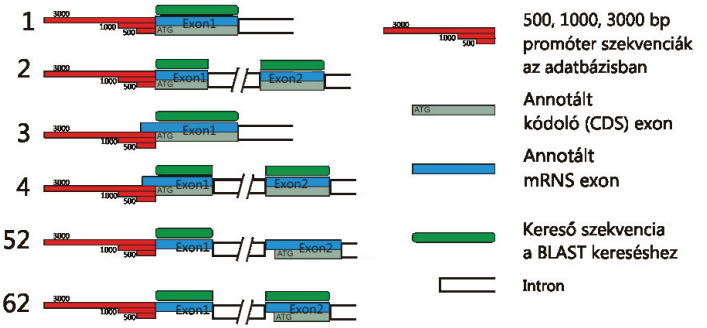
Mindkét adatbázis elkészítésénél hasonló logika szerint választottuk ki a keresőszekvenciákat, amelyekkel a felsorolt NCBI BLAST adatbázisokban és a különféle genomokban kerestünk.

A keresőszekvenciák 6 nagy csoportra oszthatók tulajdonságaik alapján. Ahol lehetséges volt kódoló exonokat használtunk fel a homológ szekvenciák kereséséhez, mivel a nem kódoló régiókkal való keresés, az e régiók kismértékű konzerváltsága miatt nem adott volna megfelelő eredményeket. A csoportok az első mRNS exon proteinkódoló szekvenciához (CDS) viszonyított startpozíciója és az első kódoló exon hossza alapján lettek megállapítva. A különböző csoportok (1, 2, 3, 4, 5*n*, 6*n*) részletes leírása a következő:

- 1. típus: a génhez tartozó mRNS és CDS első exonok startpozíciója megegyezik, az első CDS exon hosszabb, mint 50 nukleotid.
- 2. típus: a génhez tartozó mRNS és CDS első exonok startpozíciója megegyezik, az első CDS exon rövidebb, mint 50 nukleotid.
- 3. típus: a génhez tartozó mRNS első exonja 5’ irányban található az első CDS exonhoz képest és az első CDS exon hosszabb, mint 50 nukleotid.
- 4. típus: a génhez tartozó mRNS első exonja 5’ irányban található az első CDS exonhoz képest és az első CDS exon rövidebb, mint 50 nukleotid.
- 5*n* típus: a génhez tartozó mRNS több mint 1 exonja (*n* számú) 5’ irányban található az első CDS exonhoz képest és az első mRNS exon hosszabb, mint 50 nukleotid.

- 6*n* típus: a génhez tartozó mRNS több mint 1 exonja (*n* számú) 5’ irányban található az első CDS exonhoz képest és az első mRNS exon rövidebb, mint 50 nukleotid.

A BLAST keresések során az 1. és 3. típusnál az első kódoló exont, a 2. és 4. esetben pedig az első két kódoló exont használtuk fel. Az 5*n* és 6*n* típusnál az 5’ nem transzlálódó régió annotált exonjait kellett felhasználnunk a keresésekhez.



2. ábra A keresőszekvenciák típusai.

Az adatbázis készítése során felhasznált keresőszekvenciák típusai az első mRNS exon proteinkódoló szekvenciához (CDS) viszonyított startpozíciója és az első kódoló exon hossza alapján. Ha ez, mint az ábra is mutatja, a 2. exon, akkor 52 illetve 62 a típus neve, általánosságban azonban 5n és 6n néven hivatkozunk rájuk. A felhasznált keresőszekvenciákat a zöld sávok jelölik, az adatbázisban felhasznált 500, 1000 és 3000 nukleotid hosszú promóter régiókat pedig a vörös sávok.

Minden keresőszekvencia egy 8 számból álló egyedi azonosítóval rendelkezik az adatbázisban, amelyet a következő algoritmus szerint generáltunk: növények esetében az *Arabidopsis thaliana* gén TAIR (*The Arabidopsis Information Resource*) azonosítójában az ’At’ részt ’8’-ra, a ’g’-t pedig 0-ra cseréltük, tehát az At2g01060 azonosítóból 82001060 lesz. A gerinces adatbázis esetében az emberi gének azonosítójának első számjegye szintén 8, a második és harmadik a kromoszóma alapján, ahonnan származik, 01-től 22-ig terjed, az X kromoszóma esetében 24, az Y

esetében pedig 25. Az ezt következő 5 számjegy egy egyszerű sorszám, a gének sorrendje alapján a genom annotációban.

4.5. BLAST keresés

Az előzőekben részletezett 1 vagy 2 exont tartalmazó keresőszekvenciákat használtuk fel, hogy a letöltött és megszűrt szekvencia adatokból kiválasszuk a feltehetőleg ortológ szekvenciákat. Ehhez a BLAST [135] nevű programcsomag 2.2.15-ös verziójának BLASTN programját használtuk az 2. táblázatban leírt paraméterekkel. A BLAST keresések eredményeit a következőképpen szűrtük: a találat minimum 50%-os hasonlósági (*identity*) értékkel rendelkezik, és legalább 75% arányú az illesztett szekvencia hossza, a keresőszekvencia méretének figyelembevételével.

Word size	12
Gap opening penalty	1
Gap extension penalty	1
E-value	0,01

2. táblázat BLAST paraméterek az ortológ kereséshez.

A szűrt eredményekből választottuk ki a továbbiakban a feltételezhetően ortológ szekvenciákat. A következő igen egyszerű algoritmust használtuk: minden egyes fajnál ahol egynél több találatot kaptunk, a legmagasabb *BLAST score* értékkel rendelkezőt választottuk ki. Amennyiben több ugyanolyan értékű találat is volt egy adott fajból, azt választottuk, amely a legtovább terjedt 5' irányba.

Ezután összegyűjtöttük az összes 500, 1000 és 3000 nukleotid hosszú szekvencia régiót a találatoktól 5' irányba, amelyeket promóterként definiáltunk és a további munkához felhasználtunk. Amennyiben a szekvenciák nem érték el az 500, 1000 vagy 3000 nukleotid hosszt, a szükséges minimum méret az adatbázisba kerüléshez 300, 700 vagy 2000 nukleotid volt. A növényi adatbázis esetében a 300 helyett már 200 nukleotid hossz is elegendő volt az adatbázisba kerüléshez.

Ezután az adott *Arabidopsis* vagy humán génhez tartozó promótert és az ortológ gének promóter régióit az általunk generált azonosítóval ellátva egy-egy fájlba

gyűjtöttük, ortológ promóter csoportokat létrehozva, amelyek a további munka alapjául szolgáltak.

4.6. Monofiletikus csoportok készítése

A promóterek részletes elemzése előtt a csoportokban lévő szekvenciákat kisebb alcsoportokba soroltuk. Az adatbázis kezdeti verzióiban ez még nem történt meg, azonban igen hamar nyilvánvalóvá váltak azok a problémák, amelyek az irodalmi áttekintés filogenetikai módszereket felhasználó motívumkereső módszereinél már említésre kerültek. A növényi és gerinces adatbázisnál is rengeteg olyan ortológ promóter csoport volt, amelyeknél konzervált motívumokat veszítettünk egy – egy távoli rokon faj ortológ szekvenciája miatt. E probléma miatt a filogenetikai távolság alapján fokozatosan bővülő monofiletikus alcsoportokat definiáltunk. A növényi adatbázis esetén 4 ilyen alcsoportot határoztunk meg, gerinceseknél 10-et.

4.6.1. Növényi alcsoportok

A *Brassicaceae* (B) alcsoport kizárólag a *Brassicaceae* családból származó szekvenciákat tartalmaz, ami értelemszerűen tartalmazza az *Arabidopsis thaliana* referenciagenomból származó promótereket, egyéb *Arabidopsis* és *Brassica* fajok mellett.

Az *eudicotyledons* (E) alcsoport minden esetleg előforduló valódi kétszikű szekvenciát tartalmaz a *Brassicaceae* fajok mellett, mint például a fekete nyár (*Populus trichocarpa*), ricinus (*Ricinus communis*), szőlő (*Vitis vinifera*) vagy lucerna (*Medicago truncatula*).

A *Magnoliophyta* (M) alcsoport már az egyszikű szekvenciákat is tartalmazza, amennyiben azok előfordulnak az adott ortológ csoportban. Ez jellemzően a rizs (*Oryza sativa* és *Oryza japonica*) vagy kukorica (*Zea mays*) szekvenciákat jelenti, habár egyéb fajok is előfordulhatnak.

A *Viridiplantae* (V) alcsoport minden olyan zöld növényi szekvenciát is tartalmaz, amelyek nem tartoztak a 3 előző kategória egyikéhez sem.

4.6.2. Gerinces alcsoportok

A *Primates* (P) alcsoport kizárólag főemlős szekvenciákat tartalmaz a referencia humán szekvenciák mellett.

Az *Euarchontoglires* (R) alcsoportba már beletartoznak a főemlősökön kívül a rágcsálók is, mint például a patkány (*Rattus norvegicus*) vagy egér (*Mus musculus*), de a nyúlalkatúak, repülő lemurok és mókuscickányok is.

Az *Eutheria* (E) kategória minden valódi méhlepényes emlőst magába foglal, ezekben az alcsoportokban már megtalálhatók például a kutya (*Canis familiaris*), macska (*Felis catus*) vagy szarvasmarha (*Bos taurus*) szekvenciák is.

A *Theria* (H) alcsoport további, az előző kategóriába nem tartozó, de méhlepényes emlősöket is tartalmazza, ez általában az oposszum (*Monodelphis domestica*) szekvenciákat jelenti.

A *Mammalia* (M) alcsoport minden emlőst, így a kacsacsőrű emlőst (*Ornithorhynchus anatinus*) és esetleg egyéb nem méhlepényes emlősfajok szekvenciáit is összegyűjti.

Az *Amniota* (A) alcsoport minden magzatburokkal rendelkező gerincest tartalmaz, így a madarak mellett – amelyek közül a tyúk (*Gallus gallus*) szekvenciák fordulnak elő a leggyakrabban – az összes hüllő szekvenciát is.

A *Tetrapoda* (T) kategória a kétéltűek szekvenciáit is tartalmazza, egyik legjellemzőbben előforduló faj a sarkantyús karmosbéka (*Xenopus tropicalis*).

A *Teleostomi* (F) alcsoport minden csontos gerincest tartalmaz, így a halak nagyrészét is, beleértve a zebradániót (*Danio rerio*) vagy fugut (*Takifugu rubripes*) is.

A *Vertebrata* (V) alcsoport minden gerinces fajt tartalmaz, ami esetleg az előző csoportokba nem fért bele, ez néhány porcos halakból származó szekvenciát jelent.

A *Chordata* (C) alcsoport a maradék gerinchúros fajokat tartalmazza, a két zsákállat fajt (*Ciona intestinalis* és *Ciona savigny*) is beleértve.

4.7. Promóter csoportok illesztése

A promóterek konzervált régióinak definiálásához először egy többszörös szekvencia illesztésre volt szükség, amit a DIALIGN2 [136] nevű program 2.2-es verziójával végeztünk el. A DIALIGN2 egy lokális szekvencia illesztést végző program, promóterek és egyéb, csak kisebb blokkokban konzerváltságot mutató szekvenciák esetében jobban használható, mint az általánosan ismert, globális szekvencia illesztést végző programok.

4.8. Ismétlődő szekvenciák maszkolása

Annak érdekében, hogy minél pontosabb szekvencia illesztéseket kapjunk, a promóterekből kitakartuk a rövid ismétlődő szekvenciákat (*simple sequence repeat*, SSR), amelyek így nem torzították a valóban konzervált régiók illesztését. Ehhez a feladathoz a *Tandem Repeats Finder* (TRF) [137] nevű programot használtuk, a 3. táblázatban összefoglalt paraméterekkel.

Match	2
Mismatch	7
Delta	7
PM	80
PI	10
Minscore	24
Maxperiod	100

3. táblázat TRF paraméterek a rövid ismétlődő szekvenciák kereséséhez.

A TRF futtatása után a megfelelő pozíciókban kitakartuk N karakterekkel az ismétlődéseket, amit a DIALIGN2 nem illeszt, majd a szekvencia illesztések elvégzése után visszacseréltük őket a megfelelő nukleotidot jelző karakterekre.

4.9. Konzervált régiók meghatározása

A konzervált régiókat egy módosított *information content* (IC) érték alapján határoztuk meg. Az IC érték számítást a különböző szekvencia logók ábrázolásánál is használják [123], mi ennek egy módosított verzióját használtuk.

Az algoritmus a következő volt: a többszörös illesztés egy adott pozícióját nézve összeszámoltuk a szekvenciákban előforduló A, C, G és T nukleotidokat, a *gap*-eket és minden egyéb esetlegesen előforduló jelölést (X, N, U) ami nem tartozott az előző 5 kategóriába.

Az A, C, G, T gyakoriságok alapján kiszámoltuk az IC értéket a 4. egyenlet alapján:

$$4.\ IC_{seq} = log_2 N - \left(- \sum_{n=1}^{N} p_n log_2 p_n \right)$$

ahol az IC_{seq} érték a klasszikus *infomation content* érték adott pozícióra, amely 0 és 2 között mozog, N az előforduló különböző szimbólumtípusok száma (itt A, C, G és T, tehát 4), p_n pedig az adott szimbólum megfigyelt gyakorisága, ami 0 és 1 között mozoghat.

A módosított értéket az 5. egyenlet adja meg:

$$5.\ IC_{mod} = IC_{seq} N_{tot} - G_{tot} - X_{tot}$$

ahol az IC_{mod} a végleges érték, az N_{tot} az összes bázis (A, C, G és T) száma, G_{tot} az összes *gap* száma, X_{tot} pedig az összes egyéb jelölések száma egy adott pozícióban.

Ezután az illesztésben megkerestünk minden olyan régiót, amelynél az IC_{mod} érték elérte a maximálisan lehetséges legalább 80%-át. Ezek lettek az úgynevezett mag motívumok, amelyektől kiterjesztettük a konzervált régiók vizsgálatát. Pozíciónként lépkedve 5' és 3' irányba, ez az IC_{mod} érték legfeljebb 65%-ra csökkenhetett.

A konzervált régiók alapján egy konszenzus szekvenciát generáltunk a következő módszerrel és jelölésekkel: minden olyan esetben, amikor egy adott pozícióban a szekvenciák legalább 90%-ában ugyanolyan bázis volt, annak a bázisnak a jelölését használtuk. Ha valamely 2 bázis mindegyike legalább 45%-ban fordul elő, akkor az IUPAC kódnak megfelelő kettős bázisjelölést alkalmaztuk (lásd 4. táblázat), minden más esetben pedig N került a pozícióba.

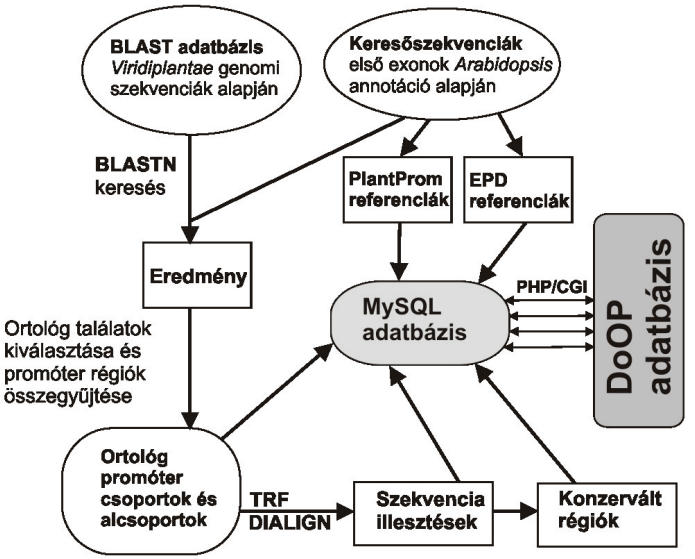
R	A vagy G
Y	C vagy T
S	G vagy C
W	A vagy T
K	G vagy T
M	A vagy C

4. táblázat Konszenzus szekvencia jelölések.

Az IC érték figyelése mellett gyakorlati tapasztalatok alapján több másik feltétel alapján is szűrtük a lehetséges evolúciósan konzervált motívumokat.

1. A minimum hossz 5 nukleotid volt, az ennél kisebbek nem adnak biológiailag értékelhető információt.
2. Adott konzervált motívumnál az illesztett szekvencia blokkban egy szekvencia sem tartalmazhatott 20%-nál több *gap*-et.
3. A konzervált motívum hosszához képest maximum 40% *gap* és N jelölésű konszenzus nukleotidot tartalmazhatott.

Az adatbázis készítés folyamatát vázlatosan a 2. ábra foglalja össze, amelyen a növényi szekció részfolyamatai láthatók, feltüntetve a fontosabb részfolyamatokat és felhasznált programokat.



3. ábra A növényi adatbázis készítésének folyamata.

A gerinces adatbázis hasonlóan készül, azonban a keresőszekvenciák a humán genomból származnak, a BLAST adatbázis az elérhető *Chordata* genomi szekvenciákat tartalmazza, a PlantProm referenciák helyett pedig a PromoSer és a DBTSS adatbázis adatait használjuk fel.

5. Eredmények

5.1. Az adatbázis generálása során felmerülő problémák

A növényi és gerinces adatbázis generálása során is előfordultak bizonyos típusú problémák, amelyek szinte minden verziónál felmerültek. Az adatbázis tartalmának bemutatása előtt kitérek ezek közül a lényegesebbekre.

5.1.1. Genom annotációval kapcsolatos problémák

Az első komolyabb probléma a referencia genomok annotációjával kapcsolatos. Az *Arabidopsis* és humán genom annotációja is folyamatosan változik az új kísérletes eredményeknek és a folyamatosan fejlődő bioinformatikai módszereknek köszönhetően. Ennek következtében minden új adatbázis verzió elkészítésekor figyelembe kell venni az annotációban történt esetleges változásokat. Ez jelentheti gének eltűnését, összevonását, több különböző génre való szétválasztását, esetleg a pontos genomi pozíciók megváltozását. A felhasználók részéről jogos elvárás emellett, hogy adott gén ugyanolyan azonosítóval szerepeljen minden verzióban. A növényi adatok és az *Arabidopsis* referencia genom esetében ez aránylag egyszerűen megoldható, mivel az At azonosító nem változik a genom annotáció különböző verziói között, így megfelelően nyomon követhetők a gének. A gerincesek esetében azonban ilyen stabil azonosító nem állt rendelkezésre, ezért minden egyes alkalommal szükséges volt egy külön BLAST keresés lefuttatása, ami alapján az új és a régi verzió génjei megfeleltethetőek voltak egymásnak.

Problémát jelentett egyes esetekben az is, hogy adott génhez tartozó CDS és mRNS annotációk alapján a gének nem rendelkeztek 5’ UTR régióval, ami valószínűleg annotációs hiba, másrészt néhány esetben a gén nem tartalmazott promóter régiót. Az 5*n* és 6*n* típusú keresőszekvenciáknál pedig kénytelenek voltunk nem kódoló exonokat felhasználni a kereséshez, habár ezek jóval kevésbé konzerváltak, mint a kódoló exonok.

5.1.2. Szekvencia adatokkal kapcsolatos problémák

A különböző szekvenáló módszerek fejlődésének köszönhetően folyamatosan növekszik a feldolgozandó teljes genomok száma és természetesen a kisebb

szekvenáló projektekből származó szekvenciák mennyisége is. A lehetséges duplikációk szűrése, a nem referenciaként használt teljes genomok feldolgozása komoly feladat volt már a tényleges BLAST keresések és azok eredményeinek elemzése előtt is. Sok esetben a nem teljes genomból származó szekvenciák nem tartalmaztak felhasználható méretű promóter régiót, ami azonban csak a keresések lefutása után derült ki.

A gerinces és növényi adatbázis esetén is fontos volt a BLAST kereséshez felhasznált szekvenciák faj szerinti szűrése is, pontosabban az, hogy beletartoznak-e a *Viridiplantae* vagy *Chordata* kategóriákba. Az NCBI adatbázis minden szekvenciához rendel egy taxonómiai azonosítót (*TaxID*), amelyből ez elvileg egyértelműen megállapítható. Azonban ez az azonosító, és sok esetben a pontos latin név, hónapról hónapra változik, nem csak faj, hanem család, vagy még magasabb szintű taxonómiai kategóriák esetén is, a rendszertani kutatások és besorolások változásainak köszönhetően. Emiatt minden adatbázis generálás alkalmával szükséges volt a szekvenciák rendszertani besorolásának vizsgálata is.

Mivel a rendelkezésre álló hardverek segítségével a gerinces adatbázis esetén csak a BLAST keresések futási ideje 4-6 hét között volt, igen lényeges volt a már említett duplikációk és a rendszertani, vagy egyéb okok miatt fel nem használható szekvenciák szűrése.

5.1.3. Ortológ-paralóg viszonyok meghatározás

Az adatbázis egyik alap feltételezése, hogy ortológ szekvenciákat hasonlítunk össze és illesztünk, azonban sok esetben a különféle gén, kromoszóma és genomduplikációk következtében nem állapíthatók meg a pontos ortológ-paralóg viszonyok. Ez a probléma gerincesek esetében is fennáll, a növényi szekció esetében azonban lényegesebb, köszönhetően a növényi genomok dinamikusabb szerkezetének. Mindezek ellenére azonban kénytelenek voltunk bizonyos elhanyagolásokat tenni, mivel a pontos ortológ-paralóg viszonyok megállapításához szükséges teljes genom illesztések, és részletes szinténia vizsgálatok nem voltak kivitelezhetőek ésszerű időhatárokon belül ekkora adatmennyiségre. Emiatt használtuk az Anyagok és módszerek részben leírt aránylag egyszerű algoritmust.

5.1.4. Motívumok meghatározása

A konzervált motívumok meghatározásánál elméletben rengeteg módszer állt rendelkezésre, azonban nagyrészük nem volt kivitelezhető ekkora adatmennyiségre, hasonlóan az ortológ-paralóg viszonyok megállapításának problémájához. A *de novo* motívumkereső algoritmusok nagy része nehezen skálázható, ezért a többszörös szekvencia illesztések mellett döntöttünk, annak ellenére, hogy így valószínűleg bizonyos mennyiségű fals negatív eredménnyel is számolni kell. A promóter régiók szerkezetének köszönhetően pedig nem az általánosan ismert és használt globális illesztést végző CLUSTALW programot használtuk, hanem több lokális és/vagy globális illesztést végző algoritmus tesztje után a DIALIGN2 mellett döntöttünk.

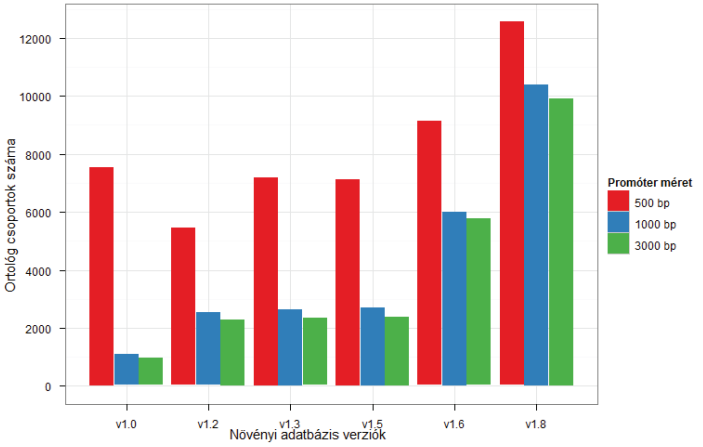
5.1.5. Adatok elérhetőségének biztosítása

Az adatok feldolgozása után, a webes keresőfelület fejlesztése közben felmerült, hogy az egyedi gének, promóterek és transzkripciós faktor kötőhelyek vizsgálata mellett sok esetben akár több 100 kísérletes eredmény vagy adatsor vizsgálatára lenne szükség, ami azonban a keresőfelületen nem, vagy csak nagyon lassan valósítható meg. Kezdetben emiatt fejlesztettük a DoOP később leírásra kerülő programozói felületét, amely később magának a keresőfelületnek a fejlesztésénél is igen hasznos volt.

5.2. Az adatbázis tartalma

5.2.1. Az ortológ csoportok és szekvenciák száma a növényi szekcióban

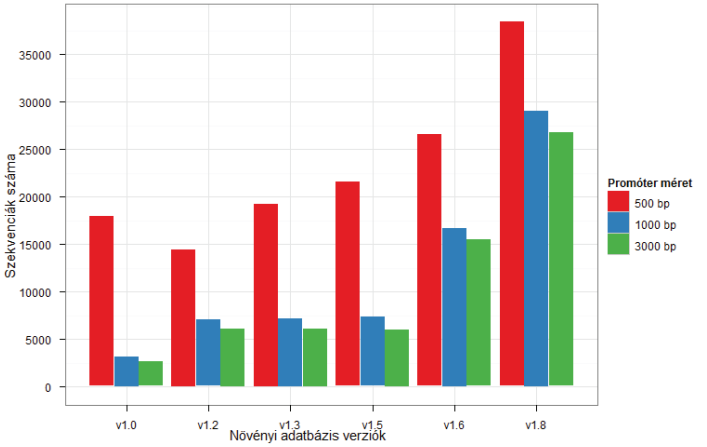
A növényi adatbázis tartalma a különböző genom szekvenáló projekteknek és egyéb kutatásoknak köszönhetően folyamatosan és jól láthatóan nőtt minden verzióban, a 3. – 4. ábra és a hozzájuk tartozó 1. – 2. online melléklet, amelyek az ortológ promóter csoportok és a szekvenciák mennyiségét ábrázolják, jól mutatják a változást. Az ábrák és a kiegészítő táblázatok tartalmazzák azon verziók alapvető adatait is, amelyek már nem érhetők el a webes felületen, a fejlesztés során történt változások miatt.



4. ábra A növényi adatbázis különböző promóter méretű és verziójú gyűjteményeiben található ortológ csoportok száma.

A növényi adatbázis esetében az 1.0, 1.2, 1.3, 1.5, 1.6 és 1.8-as verzió állt rendelkezésre, amelyek közül az 1.5, 1.6 és 1.8-as verzió érhető el a webes keresőfelületen keresztül. A 3. ábra adataiból jól látszik, hogy elsősorban az 1000 és 3000 nukleotid hosszú ortológ promóter csoportok mennyiségében történt komolyabb változás, amellett hogy az 500 nukleotid hosszú csoportok száma is növekedett, egy, az 1.2-es verziónál látható átmeneti csökkenést kivéve, amely a BLAST adatbázis összeállításával kapcsolatos módszer változásának volt köszönhető.

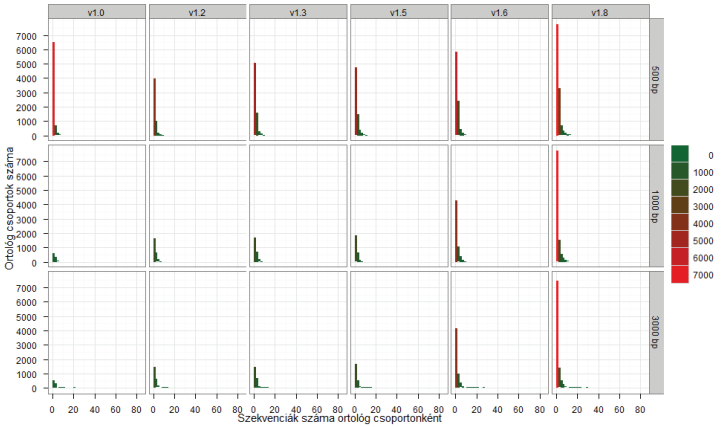
A legfontosabb változás az volt, hogy az első növényi adatbázis verzió elkészítésekor még nem állt rendelkezésre a fekete nyár, szőlő, papaya, kukorica, uborka és köles genom sem, amelyek adatai a későbbi verziókban már szerepeltek.



5. ábra A növényi adatbázis különböző promóter méretű és verziójú gyűjteményeiben található szekvenciák száma.

A szekvenciák mennyisége hasonló változást mutat a 4. ábrán az ortológ promóter csoportokéhoz. Sok esetben előfordult, hogy 200-500 nukleotid hosszúságú szekvenciák még rendelkezésre álltak egy-egy *Arabidopsis* génhez különféle forrásokból, az 1000 és 3000 nukleotidos szekcióhoz azonban már nem volt meg a szükséges minimum méret. Ezt a hiányosságot küszöbölték ki később a genom szekvenáló projektek.

A szekvenáló, különösen a kétszikű fajokat szekvenáló projekteknek köszönhető hogy az 500 nukleotidos szekciók mennyisége is nőtt, mivel egyre több olyan csoportot lehetett létrehozni, ahol a referencia *Arabidopsis* szekvencia mellett legalább 1 másik faj előfordult, ami a kezdeti verzióknál még nem volt mindig elmondható.



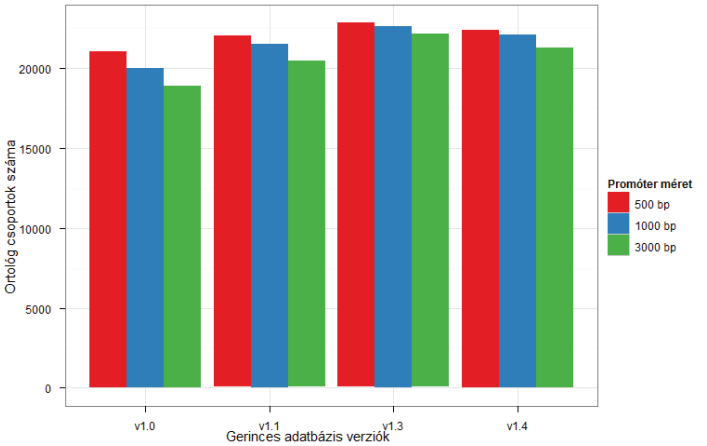
6. ábra A növényi adatbázis különböző promóter méretű és verziójú gyűjteményeiben található szekvencia szám eloszlása ortológ csoportonként.

A szekvenciák mennyiségének ortológ csoportonkénti eloszlása is jelentősen változott a verziók között. Az 5. ábra – amelynek nyers adatait a 3. online melléklet tartalmazza – hisztogramjai alapján a legszembetűnőbb az, hogy az utolsó verzióban már az 1000 és 3000 nukleotidos szekcióban is mekkora a legalább 2 szekvenciát tartalmazó csoportok száma. Emellett jól látható, hogy a kettőnél több ortológ promótert tartalmazó csoportok száma jóval kevesebb, ami a későbbi elemzések szempontjából nem volt a legideálisabb, ez remélhetőleg a jövőben az adatbázis készítés módszereinek javításával és további genomok adatainak felhasználásával javítható.

5.2.2. Az ortológ csoportok és szekvenciák száma a gerinces szekcióban

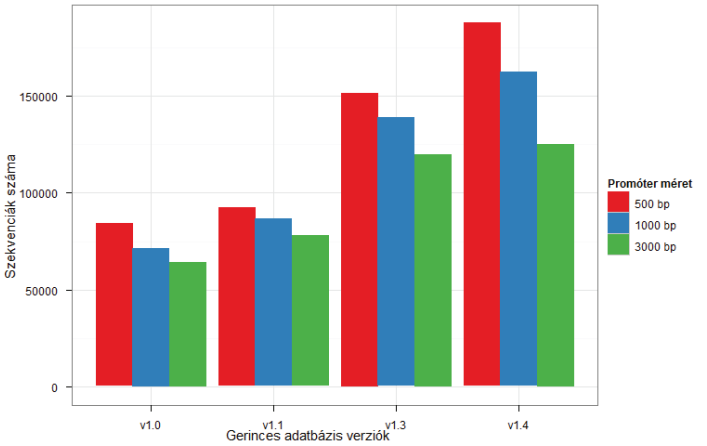
A gerinces szekció adatai kissé más változásokat mutattak a fejlesztés során a növényekhez képest. A 6. – 7. ábrák és a hozzájuk tartozó 4. és 5. online melléklet itt is tartalmazzák a weben már nem elérhető kezdeti verziók adatait az ortológ csoportok és szekvenciák számáról. Összesen 4 verzió létezik, ezek az 1.0, 1.1, 1.3 és 1.4, amelyek közül az 1.4-es érhető el publikusan. Az ortológ csoportok esetében, amit a 6. ábra mutat, a gerinceseknél nem történt lényegi változás, az 500, 1000 és 3000 nukleotidos szekciók mennyisége végig hasonló értékek körül mozog. Ennek

oka valószínűleg az, hogy már az első verzió elkészítésekor rendelkezésre állt több gerinces genom, amelyeknek köszönhetően sikerült az ortológ csoportokat legalább 2 szekvenciával létrehozni.



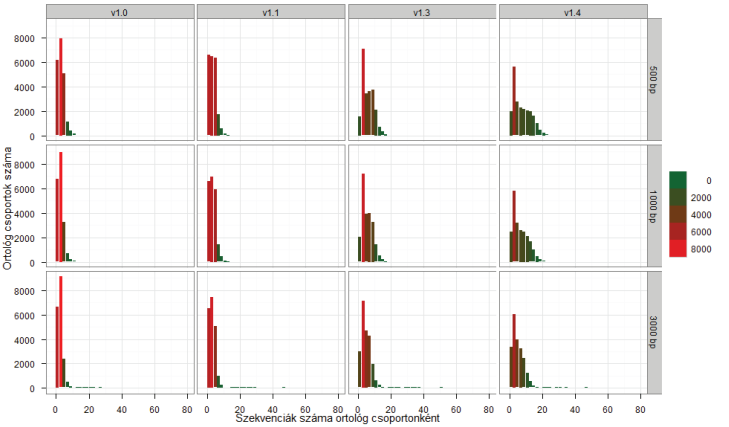
7. ábra A gerinces adatbázis különböző promóter méretű és verziójú gyűjteményeiben található ortológ csoportok száma.

A 7. ábrán látható szekvencia szám az ortológ csoportokkal ellentétben itt is jelentős emelkedést mutatott, köszönhetően a genom szekvenálásoknak, habár olyan mértékű változás nem volt tapasztalható a kezdeti mennyiséghez képest, mint a növények esetében.



8. ábra A gerinces adatbázis különböző promóter méretű és verziójú gyűjteményeiben található szekvenciák száma.

A szekvenciák ortológ csoportonkénti eloszlása azonban alapvetően különbözik a növényes szekciótól, amint az a 8. ábrán és a 6. online mellékletben látható. Nagyon kevés olyan humán gén volt, amelyhez 1 ortológ szekvenciát sem sikerült találni. Ez elsősorban a meglévő csimpánz genom szekvenciáknak volt köszönhető, mivel így szinte minden esetben legalább a csimpánz ortológ szekvencia bekerült az adatbázisba. Az újabb adatbázis verzióknál pedig jól látszik, hogy mindhárom promóter hossz esetén folyamatosan nő a legalább 5 különböző fajból származó szekvenciát tartalmazó csoportok száma.



9. ábra A gerinces adatbázis különböző promóter méretű és verziójú gyűjteményeiben található szekvencia szám eloszlása ortológ csoportonként.

5.3. A növényi adatbázis utolsó verziójának (1.8) tartalma és elemzése

5.3.1. Összes nukleotid mennyiség és az ortológ csoportok száma

Az 1.8-as növényi adatbázis nukleotid mennyiségét az 5. táblázat tartalmazza promóterek és alcsoportok szerint lebontva.

v1.8	**összesen**	***Brassicaceae***	***eudicotyledons***	***Magnoliophyta***	***Viridiplantae***
500 bp	18693437	13374195	7209104	1926309	1064388
1000 bp	28894122	19472070	12027274	2690811	1578689
3000 bp	79933237	54731869	32316138	6846016	3826918

5. táblázat Az 1.8-as verziójú növényi adatbázis összes nukleotidmennyisége promóterméret és alcsoportok szerint lebontva.

Az ortológ csoportok számát a 6. táblázat tartalmazza, az *Arabidopsis* genom annotáció alapján a kiindulási 1. (és opcionálisan 2.) exonok száma 26982 volt. Feltűnő, hogy már az 500 nukleotid hosszú promóterek esetén milyen sok exonhoz, és ez által promóter régióhoz nem találtunk ortológokat. Ennek egyik fő oka a teljesen megszekvenált növényi genomok aránylag kis száma, amelyet az egyéb szekvencia adatok, és részleges, akár csak 1-1 kromoszómára kiterjedő szekvenáló projektek sem egészítenek ki megfelelően. A másik fő ok a növényi genomok komplex dinamikája, a folyamatos poliploidizáció, esetenként csak kromoszómák, de rendszeresen teljes genomok duplikációja, amelyet aztán deléciók sorozata követ [138]. Ennek köszönhetően sokszor már közeli rokon fajok között is nehéz a pontos ortológ – paralóg viszonyok megállapítása.

v1.8	**500 bp**	**1000 bp**	**3000 bp**
Legalább 1 megfelelő méretű ortológ	12562	10407	9913
Nincs ortológ	14420	16575	17069

6. táblázat Az 1.8-as verziójú növényi adatbázis ortológ promóter csoportjainak száma promóter méret szerint lebontva.

5.3.2. Az ortológ csoportok típusai

A már részletezett 1-6*n* ortológ csoport típusok megoszlása a 7. táblázatban látható. A 3. típusból volt a legnagyobb mennyiségben, amelyeknél valószínűleg a legpontosabb az annotáció, azaz van annotált 5' nem transzlálódó régió és az első kódoló exon eléri az 50 nukleotid hosszt. A következő két legtöbb ortológ csoportot tartalmazó típus az 1 és 5*n*, ahol az annotáció esetleges hiányosságai mellett (hiányzó 5' UTR, nem kódolónak feltüntetett exonok) még mindig megfelelő volt az exon méret a kereséshez.

	500 bp	1000 bp	3000 bp
1	2974	2473	2352
3	9006	7553	7235
5*n*	582	381	326

7. táblázat táblázat Az 1.8-as verziójú növényi adatbázis ortológ promóter csoportjainak típusa promóter méret szerint lebontva.

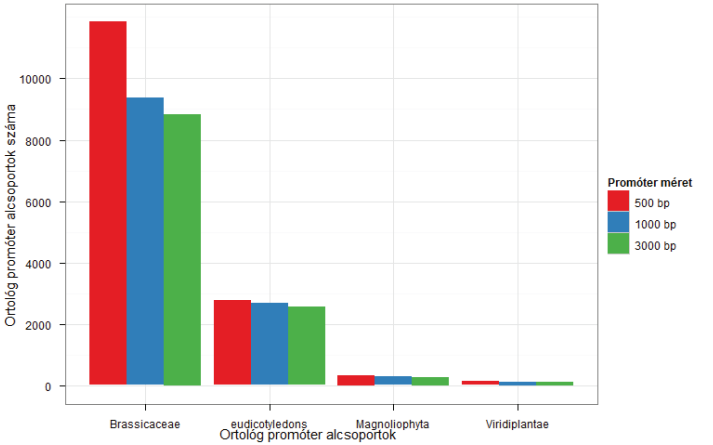
A 2, 4 és 6*n* típusok, ahol 2 exont kellett felhasználni a kereséshez a megfelelő érzékenység elérése végett, aránylag kis mennyiségben szerepeltek már az *Arabidopsis* genom annotáció feldolgozásakor is, a végleges ortológ csoportok között pedig nem fordultak elő.

5.3.3. Az ortológ promóter alcsoportok eloszlása

A promóter csoportok elemzésénél igen fontos volt a konzervált régiók meghatározása szempontjából az, hogy adott csoport milyen fajokból, milyen filogenetikai távolsággal rendelkező szekvenciákat tartalmazott. A különböző alcsoportok mennyiségének eloszlása gyors átfogó képet ad az előforduló fajokról. A 9. ábrán és a 7. online mellékletben látható, hogy legnagyobb mennyiségben a *Brassicacea* alcsoport volt megtalálható, az *Arabidopsis* referencia szekvenciához az esetek igen nagy részében csak a közvetlen taxonómiai családba tartozó fajok ortológ szekvenciáit sikerült megtalálni. Ennek oka a kevés rendelkezésre álló szekvencia mellett valószínűleg a *Brassicaceae* család genomevolúciója során történt teljes genom duplikáció [138] volt, amely megnehezítette az ortológok kiszűrését.

A valódi kétszikű szekvenciákat tartalmazó alcsoportok száma még jelentős, a másik két csoport, a *Magnoliophyta*, amely már egyszikű, és a *Viridiplantae*, amely egyéb

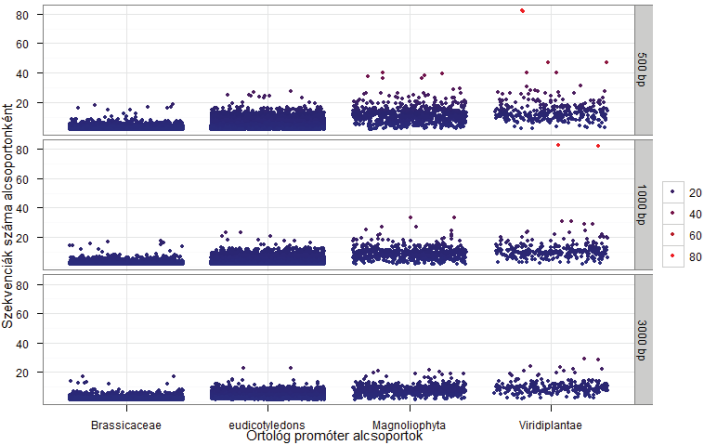
zöld növényi szekvenciákat is tartalmaz azonban majdnem elhanyagolható méretű, habár bizonyos mennyiségű konzervált régiót még így is sikerült meghatározni.



10. ábra Az 1.8-as verziójú növényi adatbázis ortológ promóter alcsoportjainak száma promóter méret szerint lebontva.

5.3.4. Szekvencia szám az ortológ alcsoportokban

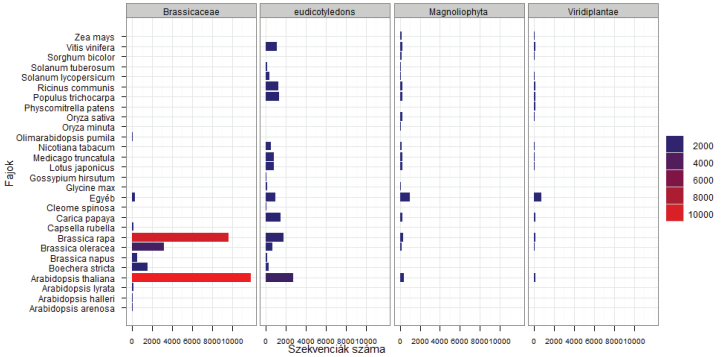
A 10. ábra és a 8. online melléklet mutatja a szekvenciák mennyiségének eloszlását a különböző promóter hosszokon és az ortológ alcsoportokon belül. Minden alcsoportnál láthatóan egy jól meghatározott tartományban mozog a szekvenciák száma, nagyon kevés olyan ortológ alcsoport van, amely kiugró értékkel rendelkezik, ezek általában régóta vizsgált, vagy központi érdeklődésre számot tartó gének, amelyeket a genomprogramoktól függetlenül több fajból megszekvenáltak.



11. ábra Az 1.8-as verziójú növényi adatbázis különböző promóter méretű gyűjteményeiben található szekvencia szám eloszlása ortológ alcsoportonként.

5.3.5. Az ortológ alcsoportokban található fontosabb fajok eloszlása

A különféle fajok eloszlásánál az 500 nukleotid hosszú promóter kollekciót elemzem részletesebben, az 1000 és 3000 nukleotid hosszú csoportokban és alcsoportokban hasonló a fajok eloszlása és aránya, amellett, hogy csökken a számuk, az esetleg nem megfelelő méretű és ez által kieső szekvenciák miatt. Az 1000 és 3000 nukleotid hosszú promóter gyűjtemények adatait is tartalmazó statisztikák a 9. online mellékletben találhatók.



12. ábra Az 1.8-as verziójú növényi adatbázis 500 nukleotid méretű promóter gyűjteményében található fontosabb fajok aránya ortológ alcsoportonként.

A különböző fajok arányai az alcsoportokban a 11. ábrán láthatók vázlatosan. A *Brassicaceae* csoportban a kis mennyiségben megjelenő *Arabidopsis* rokonfajok mellett elsősorban a *Brassicaceae* családba tartozó *Brassica rapa*, *Brassica oleracea* és a *Boechera stricta* a jelentős, mindegyikük ezernél több alcsoportban volt megtalálható. A részletes adatok a 8. táblázatban találhatók.

Faj	Szekvencia szám	Faj	Szekvencia szám
Arabidopsis thaliana	11844	*Capsella rubella*	132
Brassica rapa	9679	*Arabidopsis lyrata*	126
Brassica oleracea	3163	*Arabidopsis halleri*	70
Boechera stricta	1573	*Arabidopsis arenosa*	63
Brassica napus	523	*Olimarabidopsis pumila*	63
Egyéb fajok	306		

8. táblázat Az 1.8-as verziójú növényi adatbázis 500 nukleotid promóter méretű, *Brassicaceae* alcsoportba tartozó szekvencia gyűjteményének fontosabb fajai.

Az *eudicotyledons* csoport már a *Brassicaceae* család mellett egyéb kétszikű fajokat is tartalmaz. Ezek elsősorban már megszekvenált genommal rendelkező növények voltak, mint a papaya, fekete nyár és ricinus vagy olyan fajok, amelyek genomjának szekvenálása folyamatban volt az adatbázis készítésekor, és már rendelkezésre álltak részleges eredmények, mint például a paradicsom (*Solanum lycopersicum*).

Faj	Szekvencia szám	Faj	Szekvencia szám
Arabidopsis thaliana	2783	*Brassica oleracea*	658
Brassica rapa	1814	*Nicotiana tabacum*	509
Carica papaya	1463	*Solanum lycopersicum*	400
Populus trichocarpa	1329	*Boechera stricta*	280
Ricinus communis	1227	*Glycine max*	166
Vitis vinifera	1103	*Brassica napus*	141
Egyéb fajok	955	*Solanum tuberosum*	124
Medicago truncatula	829	*Gossypium hirsutum*	98
Lotus japonicus	803	*Cleome spinosa*	83

9. táblázat Az 1.8-as verziójú növényi adatbázis 500 nukleotid promóter méretű, *eudicotyledons* alcsoportba tartozó szekvencia gyűjteményének fontosabb fajai.

Az *eudicotyledons* csoport különböző lényegesebb fajainak gyakoriságát a 9. táblázat tartalmazza.

A *Magnoliophyta* csoport egyszikű fajokat is tartalmazott, azonban nem túl nagy számban a jelentős filogenetikai távolság és az ortológ-paralóg viszonyok megállapításának nehézségei miatt. Nagyobb mennyiségben rizs, kukorica és köles szekvenciák fordultak elő, mint az a 10. táblázat alapján is látszik.

Faj	**Szekvencia szám**	**Faj**	**Szekvencia szám**
Egyéb fajok	974	*Medicago truncatula*	195
Arabidopsis thaliana	351	*Zea mays*	177
Brassica rapa	268	*Nicotiana tabacum*	144
Vitis vinifera	252	*Sorghum bicolor*	126
Populus trichocarpa	244	*Brassica oleracea*	119
Carica papaya	240	*Solanum lycopersicum*	110
Ricinus communis	223	*Glycine max*	58
Oryza sativa	214	*Oryza minuta*	51
Lotus japonicus	195	*Solanum tuberosum*	50

10. táblázat Az 1.8-as verziójú növényi adatbázis 500 nukleotid promóter méretű, *Magnoliophyta* alcsoportba tartozó szekvencia gyűjteményének fontosabb fajai.

A *Viridiplantae* csoport esetében a *Physcomitrella patens*, a moha modellnövény volt az egyetlen, amely nagyobb mennyiségben megtalálható volt az ortológ csoportokban az eddig nem szereplő fajok közül, minden egyéb fajból minimális mennyiségű szekvencia volt (11. táblázat).

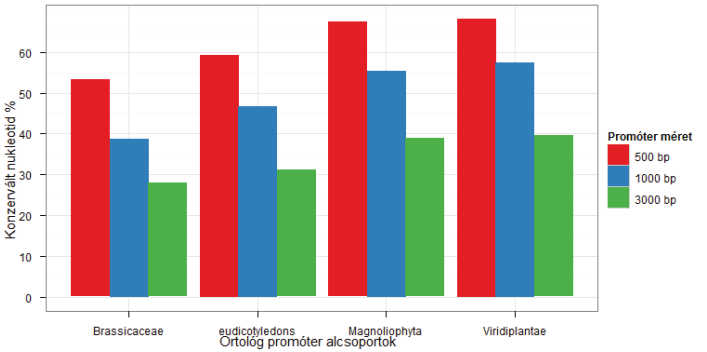
Faj	Szekvencia szám	Faj	Szekvencia szám
Egyéb fajok	726	*Medicago truncatula*	101
Arabidopsis thaliana	157	*Lotus japonicus*	100
Vitis vinifera	135	*Oryza sativa*	75
Physcomitrella patens	130	*Nicotiana tabacum*	70
Ricinus communis	127	*Solanum lycopersicum*	61
Populus trichocarpa	125	*Zea mays*	59
Brassica rapa	118	*Brassica oleracea*	51
Carica papaya	114	*Sorghum bicolor*	51

11. táblázat Az 1.8-as verziójú növényi adatbázis 500 nukleotid promóter méretű, *Viridiplantae* alcsoportba tartozó szekvencia gyűjteményének fontosabb fajai.

5.3.6. Illesztett és nem illesztett nukleotidok aránya

A DIALIGN2 program a szekvencia illesztés során külön jelöli azokat a nukleotidokat, amelyeket az illesztés során figyelembe vett a lokális illesztést végző algoritmus, és azokat, amelyeket nem. A teljes szekvencia illesztett része egy hasznos mérőszám lehet az alcsoportok és promóterek általános illesztési és konzerváltsági irányvonalainak megállapítására, habár ezek az eredmények óvatosan kezelendők. Az adatokat a 12. táblázat és a 12. ábra tartalmazza.

Az értékekből jól látszik, hogy a növényi promóter régiók első néhány 100 nukleotidnyi régiója a leginkább konzervált, valószínűleg ide összpontosul a funkcionális elemek nagy része, a promóter régió aránylag kompakt.



13. ábra Az 1.8-as verziójú növényi adatbázis különböző promóter méretű gyűjteményeinek illesztett nukleotid %-a ortológ alcsoportonként.

A másik szembetűnő változás az, hogy az egyre több fajt magába foglaló ortológ alcsoportok esetén nő az illesztett nukleotidok aránya, és ez mindhárom promóter méret esetén igaz. Ennek oka valószínűleg az, hogy a DIALIGN2 algoritmusa a szekvencia szám növekedésével egyre több nukleotidról képes megállapítani annak a többszörös illesztésben elfoglalt pontos helyét, nem pedig az egyre távolabbi fajok közötti egyre nagyobb konzerváltság, ami igen komoly ellentmondás lenne az általában elfogadott elképzelésekkel szemben.

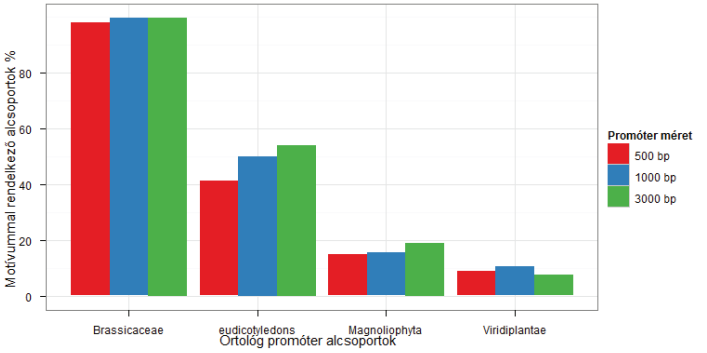
	Brassicaceae	***eudicotyledons***	***Magnoliophyta***	***Viridiplantae***
500 bp	53,21 %	59,34 %	67,35 %	68,2 %
1000 bp	38,84 %	46,78 %	55,47 %	57,47 %
3000 bp	28 %	31,1 %	38,91 %	39,64 %

12. táblázat Az 1.8-as verziójú növényi adatbázis különböző promóter méretű gyűjteményeinek illesztett nukleotid %-a ortológ alcsoportonként.

5.3.7. A konzervált motívumok száma és mérete

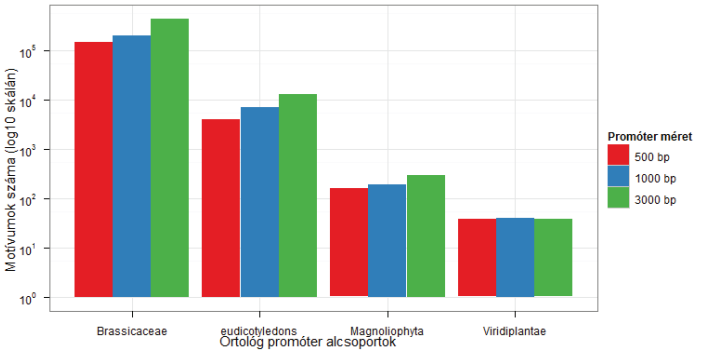
Az illesztett bázisok arányaihoz képest a szigorú szabályok szerint definiált konzervált motívumok jóval kevésbé fedik le a szekvenciákat, a *Magnoliophyta* és *Viridplantae* csoportokban drasztikus csökkenés tapasztalható a motívumok

számában, nagyrészük egyet sem tartalmaz, mint azt a 13. ábra és a 10. online melléklet is mutatja.



14. ábra Az 1.8-as verziójú növényi adatbázis különböző promóter méretű gyűjteményeinek konzervált motívummal rendelkező ortológ alcsoportjainak %-a.

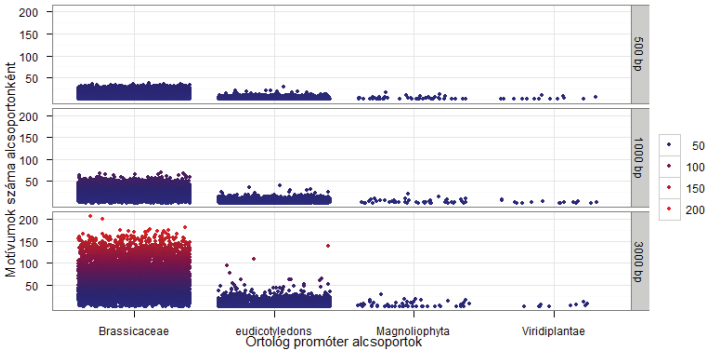
A motívumok összes számát mutatja a 14. ábra és a 11. online melléklet. A *Brassicaceae* csoportban százezres nagyságrendű a konzervált régiók száma. Ennek oka egyértelműen a nem megfelelő filogenetikai távolság, sok ortológ alcsoportban a szinte a teljes szekvencia illesztés lefedhető a motívumokkal, nincsenek jól kiugró konzervált régiók. Az *eudicotyledons*, *Magnoliophyta* és *Viridiplantae* csoportok esetében határozottan javul a helyzet, az illesztések alapján jól elkülöníthetők az egyre nagyobb filogenetikai távolságú szekvenciák közötti konzerválódott régiók.



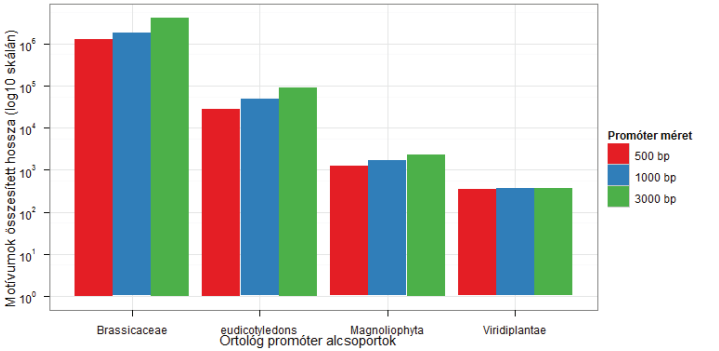
15. ábra Az 1.8-as verziójú növényi adatbázis különböző promóter méretű gyűjteményeinek összes motívumszáma.

Hasonló trendek olvashatók le a 15. ábráról és 12. online mellékletről. A *Brassicaceae* csoportban még magas az átlagos és maximálisan elért motívumszám az egyes ortológ alcsoportokban, a másik három típusnál ez gyorsan csökken. A *Brassicaceae* csoporton belül sokszor nem is beszélhetünk valódi motívumokról, 2 vagy 3 közeli rokon faj esetén az algoritmus által meghatározott motívumok lefedik majdnem a teljes szekvencia illesztést.

Ha vizsgáljuk a motívumok hosszát, és azt, hogy mekkora részét fedik le átlagosan egy-egy illesztésnek az alcsoportokon belül, hasonló eredményeket kapunk, mint amikor a motívumok számát vizsgáljuk. A 16. ábra és a 13. online melléklet a motívumok összesített hosszát mutatja, a 4 alcsoport között nagyságrendbeli eltérések vannak.



16. ábra Az 1.8-as verziójú növényi adatbázis különböző promóter méretű gyűjteményeiben található motívumszám eloszlása ortológ alcsoportonként.



17. ábra Az 1.8-as verziójú növényi adatbázis különböző promóter méretű gyűjteményeiben található motívumok összesített hossza.

Az egyes promóterek és alcsoportok átlagos motívummérete viszont nagyon hasonló volt, minden csoportnál 6 és 8 között mozgott ez az érték, ami jól egybevág az irodalmi adatokkal, miszerint a transzkripciós faktorok által felismert kötőhelyek ebben a mérettartományban mozognak, és még akkor is megtalálható egy 6-8 nukleotid hosszúságú központi, erősen konzervált régió, ha a teljes fehérje a valóságban nagyobb szakaszt fed le a DNS-ből. A részletes adatokat a 14. online melléklet tartalmazza.

5.3.8. Referencia szekvenciához viszonyított konzerváltsági arány

A konzervált motívumoknál néztük azt is, hogy a konzervált motívumok hány százalékát fedik le a szekvenciáknak. Mivel egy motívum értelemszerűen több szekvencia hasonló régióiból készül, az adott promóter és alcsoport kombináció összesített motívumhosszát osztottuk az *Arabidopsis* referencia szekvenciák összesített hosszával. A százalékos értékeket a 13. táblázat tartalmazza.

	Brassicaceae	***eudicotyledons***	***Magnoliophyta***	***Viridiplantae***
500 bp	21,88 %	2,01 %	0,71 %	0,43 %
1000 bp	18,97 %	1,77 %	0,55 %	0,26 %
3000 bp	15,68 %	1,18 %	0,27 %	0,09 %

13. táblázat Az 1.8-as verziójú növényi adatbázis különböző promóter méretű gyűjteményeinek konzerváltsági aránya.

Az arányokat vizsgálva látható, hogy az 500, 1000 és 3000 nukleotid hosszú régióknál folyamatosan csökken a konzerváltság aránya. Ez talán arra utalhat, hogy a funkcionális, evolúciósan konzerválódott részek az első néhány 100 nukleotidos régióba koncentrálódnak, a tágabb promóter környezet ilyen részeket egyre kevésbé tartalmaz.

5.4. API és MOFEXT

Az adatbázis keresőfelülete mellett egy egységes programozói felületet (API, *Application Programming Interface*) is létrehoztunk, ami a weboldal fejlesztését és a későbbi bonyolultabb parancssoros elemzések elvégzését is segítette. Mivel munkánk során igen sűrűn támaszkodtunk a Perl programozási nyelvre és annak BioPerl moduljaira, az API is ezen a nyelven készült. A Perl bioinformatikai és egyéb biológiai adatokat feldolgozó moduljai a Bio névtérben találhatók, az API pontos elnevezése Bio::DOOP lett. Ezen belül találhatók a különböző osztályok, amelyek az adatbázis tartalmának kezelésére szolgálnak.

A motívumok közötti keresésre fejlesztettük a MOFEXT programot, amely a weboldal szolgáltatásai közé is be lett építve. A program gyors, hézag nélküli illesztést végez, amely lehetővé teszi, hogy akár a több millió konszenzus szekvenciát tartalmazó gerinces motívumgyűjtemények között is belátható időn belül lefussanak a keresések. Az összehasonlítani kívánt motívumokat egységnyi darabokra bontja, majd egy hasonlósági mátrixot használ fel az azonos pozícióban található bázisok pontozására, és ezáltal egy hasonlósági érték megállapítására. A program C nyelven készült, erőforrás igénye minimális, és alapvetően rendszerfüggetlen, a forráskód alapján szinte bármely Linux, Solaris vagy OSX alapú rendszerre lefordítható és futtatható.

A Bio::DOOP különböző osztályai és rövid leírásuk a 14. táblázatban található, részletesebb, példákkal illusztrált ismertetésük és a MOFEXT pontos algoritmusa pedig Nagy Tibor disszertációjában [139].

Bio::DOOP::DOOP	DOOP API fő modul
Bio::DOOP::DBSQL	Adatbázis kapcsolat létrehozása
Bio::DOOP::Cluster	Ortológ promóter csoport reprezentálása
Bio::DOOP::ClusterSubset	Ortológ promóter alcsoport reprezentálása
Bio::DOOP::Sequence	Adott szekvencia reprezentálása
Bio::DOOP::SequenceFeature	Szekvencia tulajdonságainak reprezentálása
Bio::DOOP::Motif	Konzervált motívum reprezentálása
Bio::DOOP::Graphics::Feature	Ortológ csoport/alcsoport, szekvenciák és konzervált motívumok grafikus ábrázolása PNG formátumban
Bio::DOOP::Util::Filt	Ortológ csoportokat tartalmazó keresési eredmények szűrése
Bio::DOOP::Util::Search	Egyszerű keresőeljárások
Bio::DOOP::Util::Sort	Ortológ csoportokat tartalmazó keresési eredmények sorba rendezése
Bio::DOOP::Util::Run::Fuzznuc	Fuzznuc program futtatása
Bio::DOOP::Util::Run::GeneMerge	GeneMerge program futtatása
Bio::DOOP::Util::Run::Mofext	Mofext program futtatása

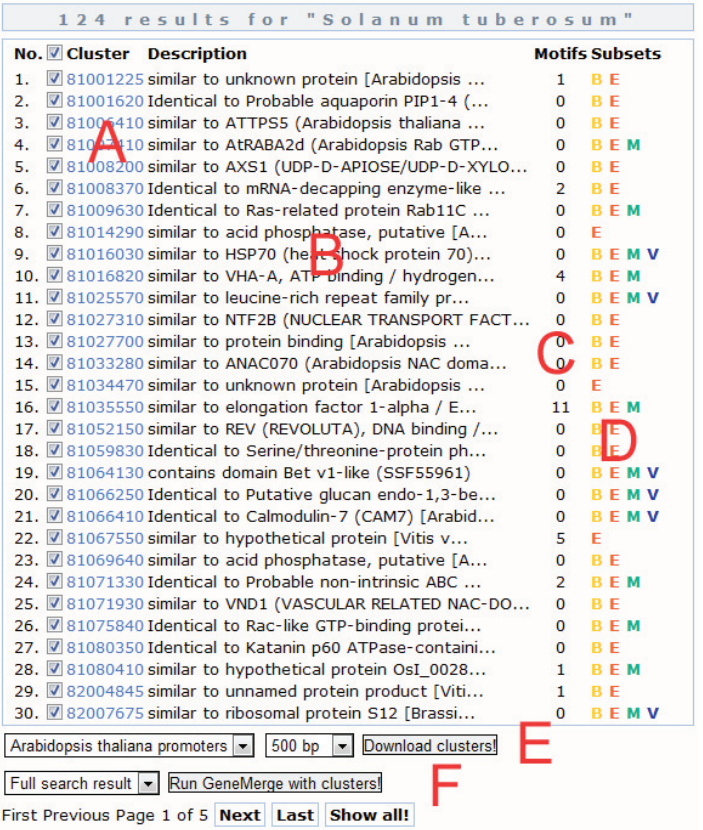
14. táblázat A DOOP API (programozási felület) osztályai és rövid leírásuk.

5.5. Keresőfelület

5.5.1. DoOP

Az adatbázishoz weben elérhető keresőfelületet is készítettünk, amellyel a leggyakoribb keresések és elemzések egyszerűen elvégezhetők. A nyitóoldalon (http://doop.abc.hu) az aktuálisan rendelkezésre álló adatbázis verziók közül lehet választani, ez gerinceseknél az 1.4-es, növényeknél pedig az 1.5, 1.6 és 1.8-as verziót jelenti. A megfelelő verzió kiválasztása után a következő oldalon lehet az adatok között keresni, különböző feltételek szerint. A lehetséges keresési módok röviden összefoglalva a következők:

- szekvencia azonosító (*Sequence ID*): adott szekvencia GI (*GenInfo Identifier*) száma – amelyet az NCBI adatbázis is használ – a külön feldolgozott teljes genomszekvenciák esetén pedig a feldolgozó programok által generált belső azonosító alapján való keresés.
- génazonosító (*Gene ID*): az NCBI *Arabidopsis thaliana* genomannotációban található génnév és szinonímái alapján való keresés.
- At azonosító (*At Number*): a TAIR (*The Arabidopsis Information Resource*) adatbázis azonosítója – amely egy átírt genomi egységre vonatkozik – alapján való keresés. Értelemszerűen csak növények esetében áll rendelkezésre.
- ENSEMBL azonosító (*ENSEMBL ID*): az ENSEMBL adatbázis humán génazonosítói alapján való keresés. Értelemszerűen csak a gerinces adatbázis esetén áll rendelkezésre.
- kulcsszó (*Keyword*): a kulcsszavas keresés az adott genom annotációban a génhez kapcsolódó bármely felhasználható mRNS vagy CDS leírás, illetve azok fehérjévé lefordított termékeinek leírásaiban keres.
- gén ontológia (*Gene Ontology*): a gén ontológia keresés 2 különböző típusú keresést végezhet. Amennyiben a kereső kifejezés 7 számot tartalmaz (pl. 0046872) vagy a GO: előtagot és ezután 7 számot (pl. GO:0046872), automatikusan a génekhez rendelt GO annotáció azonosítói közötti keresés történik. Ha bármilyen egyéb kereső kifejezés kerül beírásra, akkor az adatbázis a GO annotációk leírása közötti kulcsszavas keresést végez.
- faj (*Taxon*): lehetséges adott fajok szekvenciáit tartalmazó promóter csoportok keresése is, itt egy legördülő menüből választható ki a rendelkezésre álló fajok közül a keresett.



18. ábra A DoOP keresési eredménye.

A) Ortológ promóter csoportazonosító és hivatkozás a részletes leírásra. B) A csoport rövid leírása. C) Konzervált motívumok száma. D) A csoportban megtalálható alcsoport típusok. E) Szekvencia letöltő menü. F) GeneMerge elemző menü.

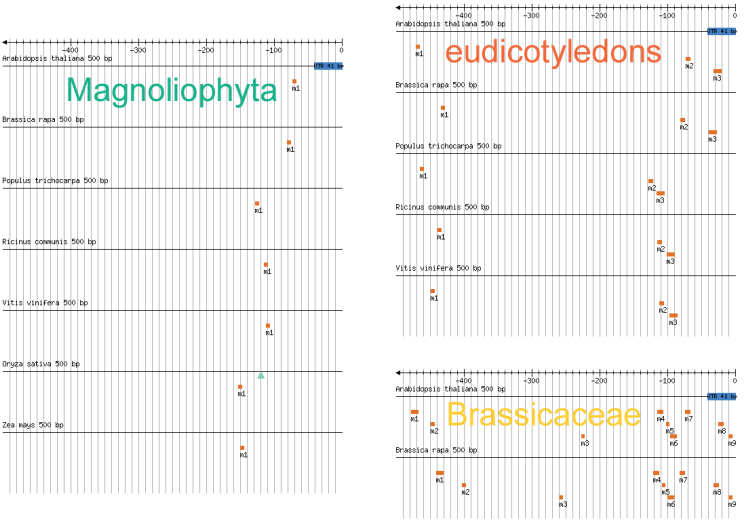
A keresési eredmények egy egységesen felépülő eredménytáblázatban jelennek meg (17. ábra). A táblázat tartalmazza az ortológ promóter csoportokra mutató hivatkozást, mindegyik csoporthoz egy rövid leírást, a legnagyobb evolúciós távolságot felölelő alcsoportjában található motívumok számát, és az összes definiálható alcsoport típusát. Emellett lehetőség van a referenciafaj (*Arabidopsis*

thaliana vagy *Homo sapiens*) promótereinek, esetleg az összes, az ortológ csoportokban található promóternek a letöltésére és egy GeneMerge [140] elemzés elvégzésére is tetszőlegesen kiválasztott promóter csoportokon.

Az eredmények részletesen megvizsgálhatók az adott ortológ promóter csoport oldalán. A promóter csoportokról megtalálható információ 3 nagy kategóriára osztható. Az első szekcióban egy részletes táblázat felsorolja a csoportban található szekvenciák azonosítóját, hosszát, a fajt, amiből származik és az alcsoportot, amibe a faj besorolható. A táblázat fölött közvetlenül két legördülő menü segítségével beállítható a vizsgálni kívánt promóter hossz kategória és alcsoport típus kombináció. Minden esetben a táblázatban kiemelt háttérszín jelzi azokat a szekvenciákat, amelyek az adott alcsoportot alkotják. A következő szekció tartalmazza a promóter csoporthoz tartozó annotációt. Ezek a következők:

- génnév vagy nevek és szinonimák
- gén típus (1 – 6*n*)
- 5' UTR hossz
- mRNS, CDS és fehérjetermék annotációi
- gén ontológia (*GeneOntology*) annotáció
- keresztreferenciák a TAIR (növények), ENSEMBL (gerincesek) és NCBI adatbázisokra
- a vizsgált promóter hosszúságú alcsoport FASTA és DIALIGN2 formátumú többszörös szekvencia illesztése

A harmadik szekció a szekvenciák és konzervált régiók grafikus megjelenítése mellett tartalmazza a motívumok részletes adatait is, azaz a motívum azonosítóját, hosszát és konszenzus szekvenciáját. A motívumok grafikus megjelenítése látható a 18. ábrán, amely 3 különböző alcsoportot tartalmaz a 81009690-es azonosítójú, az 1.8-as növényi adatbázisban megtalálható promóter gyűjtemény 500 nukleotid hosszú régióinak illesztéséből. A motívumok adatait tartalmazó táblázatról lehet továbbjutni az egyes motívumok részletes adatait bemutató oldalig. Itt a motívum konszenzus szekvenciája mellett megtalálhatók a pontos adatok a motívumot alkotó egyes szekvencia régiók pontos elhelyezkedéséről és nukleotid sorrendjéről is.



19. ábra A 81009690-es azonosítójú, az 1.8-as növényi adatbázisban megtalálható promóter gyűjtemény 500 nukleotid hosszú régióinak illesztése 3 különböző alcsoport esetén.

Jól látható, hogy az illesztésben részt vevő szekvenciák filogenetikai távolságának növekedésével (*Brassicaceae*, *eudicotyledons*, majd *Magnoliophyta*) nagymértékben csökken a konzervált régiók száma, amelyet a narancssárga négyzetek jeleznek.

5.5.2. DoOPSearch

A DoOP adatbázis keresőfelületének fejlesztése után következő lépésként egy olyan keresőfelületet is elkészítettünk, amelynek segítségével a motívumok és promóterek szekvenciáiban lehet keresni tetszőleges egyéb motívumok után. A weboldal e része a DoOPSearch (http://doopsearch.abc.hu) névre hallgat és a DoOP oldalhoz hasonlóan, a kívánt adatbázis kiválasztása után egy keresőoldalra jutunk. Ez két nagy szekcióra osztható, annak megfelelően, hogy a konzervált motívumok között, vagy a promóter régiókban kívánunk keresni.

A motívumok közötti keresést a már ismertetett, saját fejlesztésű MOFEXT program végzi. A keresés megadandó paraméterei a következők:

- konszenzus szekvencia motívum (*Pattern*): az IUPAC nevezéktant követő (lásd 1. táblázat) és minimum 5 nukleotid hosszú szekvencia
- szóméret (*Wordsize*): az illesztésnél használt mozgó ablak mérete, amelyen belül a keresőszekvencia és az adatbázis motívumai összehasonlításra kerülnek
- hasonlósági % (*Cutoff*): a kereső szekvencia és az adatbázis motívum közötti minimum hasonlóság értéke
- hasonlósági mátrix (*Matrix*): a bioinformatikai alkalmazásokban sokszor előforduló hasonlósági mátrix [141] határozza meg a kereső szekvencia és az adatbázis motívumok esetén adott nukleotid pár hasonlósági pontértékét az illesztés folyamán, amely alapján a hasonlósági % kiszámolásra kerül

A keresési eredmények egy összefoglaló táblázatban szerepelnek a DoOP adatbázis eredményeihez hasonlóan, néhány különbséggel. Az ortológ promóter csoportok és azok rövid leírása mellett az alcsoport típusa és a promóter mérete is látható, amelyben a motívum előfordult. Emellett a motívum pontszáma, amely a keresőmotívummal mutatott hasonlóságot érzékelteti, is felsorolásra kerül. Az eredményekkel szintén lefuttatható a GeneMerge elemzés és letölthetők a promóter szekvenciák, másrészt a találatok sorba rendezhetők az ortológ csoport azonosítói vagy a motívumok pontértékei szerint, és szűrhetők a promóter méret, alcsoport és a motívum pontszám alapján.

Az eredményekről továbbjutva a promóter csoport adatait tartalmazó oldalra, az egyetlen különbség a DoOP ugyanezen oldalához képest, hogy a motívumok közül a grafikus megjelenítésen és a táblázatban is kiemeltük a találatot.

A promóter szekvenciákban való kereséshez a FUZZNUC programot használtuk fel az EMBOSS programcsomagból. Ebben az esetben is több különböző paramétert szükséges megadni:

- konszenzus szekvencia motívum (*Pattern*): az IUPAC nevezéktant követő (lásd 1. táblázat) és minimum 5 nukleotid hosszú szekvencia
- hibás nukleotidok (*Mismatch*): az engedélyezett nem egyező nukleotidok száma a kereső szekvencia és a promóter szekvencia között

- komplement (*Complement*): keressen-e a program a promóter komplementer szálán vagy ne
- promóter méret (*Promoter size*): mely promóter hosszúságú gyűjteményben történjen a keresés
- promóter típus (*Promoter type*): csak a referencia genomok szekvenciáiban, vagy az összes szekvenciában történjen a keresés

Az eredménytáblázat szintén hasonlít a DoOP eredményeire, azonban az ortológ promóter csoportok és a rövid leírás mellett a promóter mérete, a DNS szál irányultsága, találatot tartalmazó szekvencia azonosítója, a találat helye és a hibás nukleotidok száma is szerepel. A megszokott GeneMerge elemzés és a szekvenciák letöltése mellett lehetőség van az ortológ csoport azonosítók és a hibás nukleotidok száma szerinti sorba rendezésre, vagy a hibás nukleotidok száma és a promóter méret szerinti szűrésre.

Adott promóter csoport oldalán pedig kiemelten szerepel a találatot tartalmazó szekvencia és a grafikus ábrázoláson is szerepel annak pontos helye.

6. Az eredmények értékelése

6.1. Összehasonlítás egyéb promóter adatbázisokkal

A DoOP adatbázis gyűjteménye nagy mennyiségű promóter szekvenciát tartalmaz a lehető legtöbb gerinces és zöld növény fajból, amelyeknek szekvenciái a készítéskor rendelkezésre álltak. Pillanatnyilag egy hasonló adatbázis érhető el, amely növényi és állati promótereket is tartalmaz, ez pedig az EPD. Nagy hátránya azonban, hogy alapvetően az irodalmi adatok alapján, manuálisan elemzett adatokat gyűjt, és ez a módszer egyre kevésbé megfelelő a szekvenáló módszerek fejlődésével exponenciálisan növekvő adatok feldolgozására. Egyetlen olyan szekcióval rendelkezik a rizs esetében, ahol a promóter szekvenciák összegyűjtése automatikus, azonban a módszer más fajokra nem lett kiterjesztve.

Az összes többi promótereket tartalmazó adatbázis fő problémája, hogy kevés fajt tartalmaz, mint például gerincesek esetében a DBTSS, OMGProm vagy PromoSer, amelyek csak néhány, általában ember, egér, patkány, esetleg egyéb emlős vagy tyúk szekvenciákkal rendelkeznek, növények esetében pedig az Athena és az Osiris, amelyekben kizárólag *Arabidopsis* és rizs promóterek érhetők el. Ezek mellett több kisebb adatbázis, mint például a HemoPDB, MPromDB vagy a TiProDB túlságosan is specializáltak, általános elemzésekre és keresésekre nem alkalmasak.

6.2. Összehasonlítás különböző transzkripciós faktor kötőhely adatbázisokkal

A különféle transzkripciós faktor kötőhely adatbázisok száma jóval nagyobb, mint a promótereket is tartalmazó gyűjteményeké, azonban hasonló típusú problémák jelentkeznek itt is, mint a promóterek esetében. Kevés az átfogó, nem csak adott fajra vagy fajcsoportra koncentráló gyűjtemény. A JASPAR adatbázis igen jó minőségű, azonban kis mennyiségű, manuálisan szűrt adatot tartalmaz. Hasonló a probléma a TRRD adatbázissal is. Az oRegAnno átfogóbb, azonban így is hiányzik sok fontos faj információja. A TRANSFAC adatbázis tekinthető az egyik legkomolyabb, legtöbb fajt felölelő adatbázisnak, részletes információkat tartalmaz rengeteg transzkripciós faktorról, promóter régióról, kötőhelyekről és a kapcsolódó

információkról, azonban üzleti alapon működik, a magas éves előfizetési díj sok esetben limitáló tényező lehet.

Az egyéb adatbázisok esetén hasonló problémák merülnek fel, mint a promóter gyűjteményeknél. A gerincesek esetében az ABS vagy a TRED igen kevés adatot tartalmaz, a cisRED pedig mindössze 4 fajt vesz figyelembe. A növényeknél az AGRIS, AthaMap és DATF az *Arabidopsis* adatokra koncentrálnak, a PLACE vagy PlantCARE hosszabb ideje nem frissül, és csak irodalmi adatokat tartalmaz. A PlantTFDB és PlnTFDB a legátfogóbb és használhatóbb adatbázisok növényi területen, azonban értelemszerűen hiányoznak belőlük a gerinces információk.

6.3. A filogenetikai lábnyom módszer előnyei és hátrányai

Az elmúlt évtizedben rengeteg promóterelemző, transzkripciós faktor kötőhelyeket előrejelző algoritmus és program alapja volt a filogenetikai lábnyom alkalmazása. A módszer alapvető és szükséges feltétele, hogy a gének promóterében a funkcióval bíró régiók evolúciósan konzerválódjanak a körülöttük lévő szekvenciához képest, amelyekre szelekciós nyomás nem vagy csak jóval kevésbé hat. A módszer sok esetben eredményesen használható, azonban bizonyos problémák is felbukkannak alkalmazása során.

A legnagyobb problémát a különféle teljes vagy részleges genomduplikációk okozzák. A gerincesek esetében aránylag kis mennyiségű teljes genomduplikációval kell számolni [142], a gének rokonsági viszonyai könnyebben felderíthetőek, habár a halak vonalán bekövetkezett extra genomduplikáció okozhat problémákat. A növények esetében sokkal bonyolultabb a helyzet [138], úgy tűnik, hogy a rendszeresen bekövetkező részleges vagy teljes genomduplikációk az állatokétól radikálisan eltérő életmódjuknak és túlélési stratégiájuknak köszönhető [143]. Ezek a genomduplikációk később komoly problémát jelentenek az ortológ-paralóg kapcsolatok pontos felderítésében.

A DoOP adatbázis ortológ csoportjainak számát és azok tartalmát vizsgálva is nyilvánvalóak ezek a trendek, még abban az esetben is, ha figyelembe vesszük, hogy a növények sok esetben csak a kutatások másodvonalában szerepelnek, jóval kevesebb teljes vagy részleges genomszekvencia állt rendelkezésre az adatbázis

elkészítésekor. A gerincesek esetében nagyon kevés olyan humán referencia gén volt, amelyhez ne találtunk volna legalább 1 feltételezett ortológot, és még azon alcsoportok száma is 1000 fölött volt, amelyek valamilyen hal promóter régiót tartalmaztak. A növényi szekció esetében jól látható, hogy a *Brassicaceae* alcsoport aránylag sok szekvenciát tartalmazó gyűjteményei mellett az eudicotyledons csoport tartalma határozottan csökkent, az egyszikűeket is tartalmazó *Magnoliophyta*, és a teljes zöld növényi vonalat felölelő *Viridiplantae* csoport pedig nagyságrendekkel kevesebb szekvenciát tartalmaz. Mindez annak ellenére, hogy a fekete nyár, ricinus, szőlő, rizs és kukorica genom szekvenciái is részlegesen vagy teljesen rendelkezésre álltak.

A filogenetikai lábnyom módszer másik lényeges problémája az ortológ-paralóg viszonyok megállapítása után az összehasonlított szekvenciák közötti filogenetikai távolság. A témában írt egyik első átfogó elemzés még minden felhasználható szekvenciaadatot figyelembe vett az elemzéskor, humán, egér, patkány és kutya szekvenciákat is felhasználva az elemzéshez [144]. A későbbiekben nyilvánvalóvá vált, hogy a promóter régiókban konzervált kötőhelyek keresése során az éppen felhasznált ortológ fajok segítségével finomhangolható a módszer [145] [146]. Ezen megfontolások alapján határoztuk meg a különböző alcsoportokat az adatbázisban, habár kezdetben még csak egy szekvenciaillesztést és motívumkeresést végeztünk, az összes szekvenciaadatot figyelembe véve.

Összességében azonban a módszer és az erre épülő adatbázis – különösen a növényi szekció – hiányosságai ellenére jól használható, a feldolgozott promóter régiók mennyiségét és a fajok számát tekintve a legátfogóbbak között van.

6.4. Növényi promóter csoportok tartalma és konzerváltsága

Az eredmények szekcióban részletesen bemutatott növényi szekció utolsó verziójának eredményeiből látszik, és már a filogenetikai lábnyom módszer hatékonyságát is tárgyaló előző rész is említi, hogy növények esetében igen komoly probléma a gének ortológ-paralóg viszonyának megállapítása. Az adatok alapján talán érdemes lenne akár családszinten elkülöníteni a fajokat, de legalább az egyszikű és kétszikű csoportok különálló elemzése elkerülhetetlennek látszik a

jövőben. Az adatbázis készítése és a konzervált régiók elemzése során felmerült az is, hogy mivel a különböző motívumkereső és elemző módszerek alapvetően gerinces fajokat (ember, egér, patkány, tyúk, zebrahal) használnak modellként, és a bioinformatikai eredmények ellenőrzése is általában e modellfajokon történik, a növények esetében nem biztos, hogy minden módszer megfelelően használható. Valószínűleg e módszerek is némi finomhangolást fognak igényelni, hogy megfelelően alkalmazhatóak legyenek növények esetében.

6.5. Az adatbázis gyakorlati felhasználása

Az adatbázist már első, interneten is elérhető verziójától kezdve felhasználták különböző kutatócsoportok egyes gének és géncsoportok szabályozásának és expressziójának vizsgálatához. Annak ellenére, hogy a gerinces adatbázis jóval átfogóbb és több hasznos adatot tartalmaz, az említett kutatásokat leíró hivatkozások nagyrésze növényi témákkal foglalkozik, amely azt mutatja, hogy igen nagy szükség van egy növényi promótereket és feltételezett transzkripciós faktorokat tartalmazó jó minőségű adatbázisra, hiszen a DoOP növényi szekciója még az aktuálisan módszertanilag nem teljesen kiforrot állapotában is hiánypótló.

Egy Arabidopsis stresszválasszal foglalkozó cikk [147] a válaszban szerepet játszó gének promótereiben lévő feltételezett transzkripciós faktor kötőhelyek adatait használja fel egy neurális hálón alapuló modell felépítésére, amely képes lehet a válaszban szerepet játszó, de még nem azonosított gének felderítésére.

Egy másik tanulmány [148] *Arabidopsis*, fekete nyár és eukaliptusz fajok ortológ cellulóz szintáz génjeinek promótereiben vizsgálja az evolúciósan konzerválódott szabályozó elemeket, egy harmadik esetben [149] pedig hosszú szénláncú zsírsavak szintézisében szerepet játszó gének promótereinek információit használták fel az adatbázisból.

A további, említésre nem kerülő cikkek alapján, és a felhasználói visszajelzésekből is látszik, hogy elsősorban a növényi szekció fejlesztésére van igény, és lesz szükség a jövőben.

7. Összefoglalás

Munkánk során kifejlesztettünk egy módszert, amelynek segítségével egy növényi és egy gerinces referenciagenom – *Arabidopsis thaliana* és *Homo sapiens* – alapján, a referenciafaj első exonjainak segítségével összegyűjtöttük a feltételezett ortológ gének első exonját a rendelkezésre álló rokon fajok szekvenciáiból. Erre alapozva meghatároztuk és összegyűjtöttük az adott ortológ génhez tartozó 500, 1000 és 3000 nukleotid hosszúságú promóterrégiókat. Az ortológ promótercsoportokat kisebb, jól definiálható, monofiletikus kategóriákra osztottuk a referencia fajhoz viszonyított filogenetikai távolság alapján. A növényi ortológ csoportok esetében 4 ilyen alcsoportot definiáltunk, a gerincesek esetében pedig 10-et.

A promóter alcsoportokban ezek után meghatároztuk a különböző evolúciósan konzervált régiókat. A motívumok definiálásának alapja egy lokális szekvenciaillesztés volt, amihez a DIALIGN2 nevű programot használtuk. A program által készített szekvenciaillesztésben egy úgynevezett „*information content*" tartalmat számolva minden oszlopra, majd az esetleges *gap*-eknek megfelelően az IC értéket súlyozva egy mérőszámot kaptunk, amely alapján a konzervált régiókat, másnéven motívumokat összegyűjtöttük.

A promótercsoportok, motívumok, és a hozzájuk kapcsolódó annotáció kereséséhez, további elemzéséhez több eszközt fejlesztettünk. A motívumok kereséséhez készült a MOFEXT nevű program, amely nagy sebességgel képes keresni a rövid motívumok között. Az összes adat publikusan elérhető a http://doop.abc.hu (DoOP, Database of Orthologous Promoters) és http://doopsearch.abc.hu (DoOPSearch) oldalakon, ahol különböző keresésekre is lehetőség van. A két weboldal és keresőfelület hátterében egy MySQL adatbázis és egy CGI/Perl scriptekből álló rendszer működik.

A webes felületen lehetőség van a gének különböző azonosítói, név vagy kulcsszó, esetleg Gene Ontology annotáció alapján való keresésére. A motívumok között és a promóter régiók szekvenciájában a MOFEXT és az EMBOSS programcsomag FUZZNUC programjával kereshetünk, az eredmények pedig több opció szerint szűrhetők, rendezhetők, és egy a GeneMerge nevű programra épülő Gene Ontology elemzésre is lehetőség van.

8. Summary

During our work we developed a new method, based on the plant *Arabidopsis thaliana* and the chordate *Homo sapiens* genomes, which collects putative orthologous first exons with the help of the first exons of the mentioned two reference genomes. Based on the data, we defined the putative orthologous promoter regions as 500, 1000 and 3000 nucleotides long regions upstream of the transcription start site. We also divided these orthologous promoter clusters to smaller monophyletic subsets, using the phylogenetic distance from the reference species. We defined 4 of these subsets in the plant orthologous promoter collections, and 10 in the chordate section.

We also determined the various evolutionary conserved regions in the subsets. The basis of this was a local sequence alignment, carried out with the program DIALIGN2. In the DIALIGN2 multiple sequence alignment, we assigned an infomation content value to each column, which was also weighted with the number of gaps and possibly unknown nucleotides. Based on this information content score we were able to analyze and collect the evolutionary conserved sequence regions or motifs.

We designed multiple tools for the search and analysis of the collected promoter regions, motifs and their annotation. First, the MOFEXT program is used for the fast indexing and search of the motifs. Also all of the data is publicly available at the http://doop.abc.hu (DoOP, Database of Orthologous Promoters) and http://doopsearch.abc.hu (DoOPSearch) addresses. A MySQL database and a CGI/Perl script system is running behind the search interfaces, complemented by an API, used in large scale analysis of promoter data.

It is possible to search for different genes based on their various identifiers, names, keywords, or their Gene Ontology annotation. Search among the motifs and promoters is carried out with the MOFEXT and the FUZZNUC program, respectively. The results might be sorted or filtered differently, and a Gene Ontology analysis is also available, based on the GeneMerge software.

9. Irodalomjegyzék

[1] J. E. F. Butler and J. T. Kadonaga, "The RNA polymerase II core promoter: a key component in the regulation of gene expression," *Genes & Development*, vol. 16, no. 20, pp. 2583-2592, 2002.

[2] G. A. Maston, S. K. Evans, and M. R. Green, "Transcriptional regulatory elements in the human genome," *Annual Review of Genomics and Human Genetics*, vol. 7, pp. 29-59, Jan. 2006.

[3] L. Tora, "A unified nomenclature for TATA box binding protein (TBP)-associated factors (TAFs) involved in RNA polymerase II transcription," *Genes & Development*, vol. 16, no. 6, pp. 673-675, Mar. 2002.

[4] F. Müller and L. Tora, "The multicoloured world of promoter recognition complexes," *The EMBO Journal*, vol. 23, no. 1, pp. 2-8, 2004.

[5] S. T. Smale and J. T. Kadonaga, "The RNA polymerase II core promoter," *Annual Review of Biochemistry*, vol. 72, pp. 449-479, Jan. 2003.

[6] S. T. Smale, "Core promoters: active contributors to combinatorial gene regulation," *Genes & Development*, vol. 15, no. 19, pp. 2503-2508, Oct. 2001.

[7] T. Juven-Gershon and J. T. Kadonaga, "Regulation of gene expression via the core promoter and the basal transcriptional machinery," *Developmental Biology*, vol. 339, no. 2, pp. 225-229, Mar. 2010.

[8] K. Florquin, Y. Saeys, S. Degroeve, P. Rouzé, and Y. Van De Peer, "Large-scale structural analysis of the core promoter in mammalian and plant genomes," *Nucleic Acids Research*, vol. 33, no. 13, pp. 4255-4264, Jan. 2005.

[9] M. L. Goldberg, "Sequence analysis of Drosophila histone genes," Stanford University, 1979.

[10] R. Breathnach and P. Chambon, “Organization and expression of eucaryotic split genes coding for proteins,” *Annual Review of Biochemistry*, vol. 50, pp. 349–383, 1981.

[11] T. Matsui, J. Segall, P. A. Weil, and R. G. Roeder, “Multiple factors required for accurate initiation of transcription by purified RNA polymerase II,” *The Journal of Biological Chemistry*, vol. 255, no. 24, pp. 11992-11996, Dec. 1980.

[12] S. K. Burley and R. G. Roeder, “Biochemistry and Structural Biology of Transcription Factor IID (TFIID),” *Annual Review of Biochemistry*, vol. 65, pp. 769-799, 1996.

[13] S. Buratowski, S. Hahn, P. A. Sharp, and L. Guarante, “Function of a yeast TATA element-binding protein in a mammalian transcription system,” *Nature*, vol. 334, no. 6177, pp. 37-42, 1988.

[14] J. M. Wong and E. Bateman, “TBP-DNA interactions in the minor groove discriminate between A:T and T:A base pairs,” *Nucleic Acids Research*, vol. 22, no. 10, pp. 1890-1896, May. 1994.

[15] V. L. Singer, C. R. Wobbe, and K. Struhl, “A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation,” *Genes & Development*, vol. 4, no. 4, pp. 636-645, Apr. 1990.

[16] P. Bucher, “Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences,” *Journal of Molecular Biology*, vol. 212, no. 4, pp. 563-578, Apr. 1990.

[17] The ENCODE Project Consortium, “Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project,” *Nature*, vol. 447, no. 7146, pp. 799-816, 2007.

[18] S. J. Cooper, N. D. Trinklein, E. D. Anton, L. Nguyen, and R. M. Myers, “Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome,” *Genome Research*, vol. 16, no. 12, pp. 1-10, 2006.

[19] P. C. FitzGerald, A. Shlyakhtenko, A. A. Mir, and C. Vinson, "Clustering of DNA sequences in human promoters," *Genome Research*, vol. 14, no. 8, pp. 1562-1574, 2004.

[20] Y. Y. Yamamoto, T. Yoshitsugu, T. Sakurai, M. Seki, K. Shinozaki, and J. Obokata, "Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis," *The Plant Journal*, vol. 60, no. 2, pp. 350-362, Oct. 2009.

[21] P. Civán and M. Svec, "Genome-wide analysis of rice (Oryza sativa L. subsp. japonica) TATA box and Y Patch promoter elements," *Genome*, vol. 52, no. 3, pp. 294-297, Mar. 2009.

[22] J. Corden, B. Wasylyk, A. Buchwalder, P. Sassone-Corsi, C. Kedinger, and P. Chambon, "Promoter sequences of eukaryotic protein-coding genes," *Science*, vol. 209, no. 4463, pp. 1406-1414, 1980.

[23] R. Grosschedl and M. Birnstiel, "Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 3, pp. 1432-1436, 1980.

[24] A. O'Shea-Greenfield and S. T. Smale, "Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription," *The Journal of Biological Chemistry*, vol. 267, no. 2, pp. 1391-1402, Mar. 1992.

[25] J. Colgan and J. L. Manley, "Cooperation between core promoter elements influences transcriptional activity in vivo," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 6, pp. 1955-1959, Mar. 1995.

[26] J. Kaufmann and S. T. Smale, "Direct recognition of initiator elements by a component of the transcription factor IID complex," *Genes & Development*, vol. 8, no. 7, pp. 821-829, Apr. 1994.

[27] G. E. Chalkley and C. P. Verrijzer, “DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator,” *The EMBO Journal*, vol. 18, no. 17, pp. 4835-4845, Sep. 1999.

[28] P. Bucher and E. N. Trifonov, “Compilation and analysis of eukaryotic POL II promoter sequences,” *Nucleic Acids Research*, vol. 14, no. 24, pp. 10009-10026, 1986.

[29] R. Javahery, A. Khachi, K. Lo, B. Zenzie-Gregory, and S. T. Smale, “DNA sequence requirements for transcriptional initiator activity in mammalian cells,” *Molecular and Cellular Biology*, vol. 14, no. 1, pp. 116-127, Jan. 1994.

[30] K. Lo and S. T. Smale, “Generality of a functional initiator consensus sequence,” *Gene*, vol. 182, no. 1-2, pp. 13-22, 1996.

[31] U. Ohler, G.-chun Liao, H. Niemann, and G. M. Rubin, “Computational analysis of core promoters in the Drosophila genome,” *Genome Biology*, vol. 3, no. 12, pp. research0087.1-0087.12, Jan. 2002.

[32] G. Yarden, R. Elfakess, K. Gazit, and R. Dikstein, “Characterization of sINR, a strict version of the Initiator core promoter element,” *Nucleic Acids Research*, vol. 37, no. 13, pp. 4234-4246, Jul. 2009.

[33] Y. Suzuki et al., “Identification and characterization of the potential promoter regions of 1031 kinds of human genes,” *Genome Research*, vol. 11, no. 5, pp. 677-684, May. 2001.

[34] C. Yang, E. Bolotin, T. Jiang, F. M. Sladek, and E. Martinez, “Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters,” *Gene*, vol. 389, no. 1, pp. 52-65, 2007.

[35] Y. Y. Yamamoto et al., “Identification of plant promoter constituents by analysis of local distribution of short sequences,” *BMC Genomics*, vol. 8, p. 67, 2007.

[36] T. W. Burke and J. T. Kadonaga, “The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila,” *Genes & Development*, vol. 11, no. 22, pp. 3020-3031, Nov. 1997.

[37] A. K. Kutach and J. T. Kadonaga, “The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters,” *Molecular and Cellular Biology*, vol. 20, no. 13, pp. 4754-4764, Jul. 2000.

[38] V. X. Jin, G. a C. Singer, F. J. Agosto-Pérez, S. Liyanarachchi, and R. V. Davuluri, “Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs,” *BMC Bioinformatics*, vol. 7, p. 114, Jan. 2006.

[39] D. B. Nikolov et al., “Crystal structure of a TFIIB–TBP–TATA-element ternary complex,” *Nature*, vol. 377, no. 6545, pp. 119-128, 1995.

[40] T. Lagrange, a N. Kapanidis, H. Tang, D. Reinberg, and R. H. Ebright, “New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB,” *Genes & Development*, vol. 12, no. 1, pp. 34-44, Jan. 1998.

[41] W. Deng and S. G. E. Roberts, “A core promoter element downstream of the TATA box that is recognized by TFIIB,” *Genes & Development*, vol. 19, no. 20, pp. 2418-2423, 2005.

[42] N. I. Gershenzon and I. P. Ioshikhes, “Synergy of human Pol II core promoter elements revealed by statistical sequence analysis,” *Bioinformatics*, vol. 21, no. 8, pp. 1295-1300, 2005.

[43] C. Y. Lim, B. Santoso, T. Boulay, E. Dong, U. Ohler, and J. T. Kadonaga, “The MTE, a new core promoter element for transcription by RNA polymerase II,” *Genes & Development*, vol. 18, no. 13, pp. 1606-1617, Jul. 2004.

[44] B. A. Lewis, T.-K. Kim, and S. H. Orkin, “A downstream element in the human beta-globin promoter: evidence of extended sequence-specific

transcription factor IID contacts," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 13, pp. 7172-7177, Jun. 2000.

[45] S. P. Persengiev, X. Zhu, B. L. Dixit, G. A. Maston, E. L. W. Kittler, and M. R. Green, "TRF3, a TATA-box-binding protein-related factor, is vertebrate-specific and widely expressed," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 25, pp. 14887-14891, 2003.

[46] F. Hirose, M. Yamaguchi, H. Handa, Y. Inomata, and A. Matsukage, "Novel 8-base pair sequence (Drosophila DNA replication-related element) and specific binding factor involved in the expression of Drosophila genes for DNA polymerase alpha and proliferating cell nuclear antigen," *The Journal of Biological Chemistry*, vol. 268, no. 3, pp. 2092-2099, Jan. 1993.

[47] Y. Tokusumi, Y. Ma, X. Song, R. H. Jacobson, and S. Takada, "The new core promoter element XCPE1 (X Core Promoter Element 1) directs activator-, mediator-, and TATA-binding protein-dependent but TFIID-independent RNA polymerase II transcription from TATA-less promoters," *Molecular and Cellular Biology*, vol. 27, no. 5, pp. 1844-1858, Mar. 2007.

[48] N. Hernandez, "Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription," *The Journal of Biological Chemistry*, vol. 276, no. 29, pp. 26733-26736, Jul. 2001.

[49] Y. Y. Yamamoto, Y. Yoshioka, M. Hyakumachi, and J. Obokata, "Characteristics of Core Promoter Types with respect to Gene Structure and Expression in Arabidopsis thaliana," *DNA Research*, vol. 18, no. 5, pp. 1-10, Jul. 2011.

[50] A. P. Bird and M. H. Taggart, "Variable patterns of total DNA and rDNA methylation in animals," *Nucleic Acids Research*, vol. 8, no. 7, pp. 1485-1497, Jan. 1980.

[51] A. P. Bird, “CpG islands as gene markers in the vertebrate nucleus,” *Trends in Genetics*, vol. 3, no. 12, pp. 342–347, 1987.

[52] M. C. Blake, R. C. Jambou, A. G. Swick, J. W. Kahn, and J. C. Azizkhan, “Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter,” *Molecular and Cellular Biology*, vol. 10, no. 12, pp. 6632-6641, Dec. 1990.

[53] R. Yamashita, Y. Suzuki, S. Sugano, and K. Nakai, “Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity,” *Gene*, vol. 350, no. 2, pp. 129-136, 2005.

[54] Y. Y. Yamamoto, H. Ichida, T. Abe, Y. Suzuki, S. Sugano, and J. Obokata, “Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis,” *Nucleic Acids Research*, vol. 35, no. 18, pp. 6219-6226, Jan. 2007.

[55] A. R. Reineke, E. Bornberg-Bauer, and J. Gu, “Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes,” *Nucleic Acids Research*, vol. 39, no. 14, pp. 6029-6043, Apr. 2011.

[56] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, “A census of human transcription factors: function, expression and evolution,” *Nature Reviews Genetics*, vol. 10, no. 4, pp. 252-263, 2009.

[57] N. M. Luscombe, S. E. Austin, H. M. Berman, and J. M. Thornton, “An overview of the structures of protein-DNA complexes,” *Genome Biology*, vol. 1, no. 1, pp. reviews001.1-001.37, Jan. 2000.

[58] G. B. Fogel et al., “A statistical analysis of the TRANSFAC database,” *BioSystems*, vol. 81, no. 2, pp. 137-154, 2005.

[59] S. C. J. Parker, L. Hansen, H. O. Abaan, T. D. Tullius, and E. H. Margulies, “Local DNA topography correlates with functional noncoding regions of the human genome,” *Science*, vol. 324, no. 5925, pp. 389-392, Apr. 2009.

[60] B. Li, M. Carey, and J. L. Workman, “The role of chromatin during transcription,” *Cell*, vol. 128, no. 4, pp. 707-719, 2007.

[61] D. a Jackson, a B. Hassan, R. J. Errington, and P. R. Cook, “Visualization of focal sites of transcription within human nuclei,” *The EMBO Journal*, vol. 12, no. 3, pp. 1059-1065, Mar. 1993.

[62] T. Cremer and C. Cremer, “Chromosome territories, nuclear architecture and gene regulation in mammalian cells,” *Nature Reviews Genetics*, vol. 2, no. 4, pp. 292-301, Apr. 2001.

[63] D. Papatsenko and M. Levine, “Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the Drosophila embryo,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 14, pp. 4966-4971, Apr. 2005.

[64] C. Geserick, H.-A. Meyer, and B. Haendler, “The role of DNA response elements as allosteric modulators of steroid receptor function,” *Molecular and Cellular Endocrinology*, vol. 236, no. 1-2, pp. 1-7, May. 2005.

[65] M. B. Harris, J. Mostecki, and P. B. Rothman, “Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function,” *The Journal of Biological Chemistry*, vol. 280, no. 13, pp. 13114-13121, Apr. 2005.

[66] L. Li, S. He, J.-min Sun, and J. R. Davie, “Gene regulation by Sp1 and Sp3,” *Biochemistry and Cell Biology*, vol. 82, no. 4, pp. 460-471, 2004.

[67] L. Srinivasan and M. L. Atchison, “YY1 DNA binding and PcG recruitment requires CtBP,” *Genes & Development*, vol. 18, no. 215, pp. 2596-2601, 2004.

[68] L. Chen and J. Widom, “Mechanism of transcriptional silencing in yeast,” *Cell*, vol. 120, no. 1, pp. 37-48, Jan. 2005.

[69] F. Recillas-Targa et al., “Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities,” *Proceedings of the*

National Academy of Sciences of the United States of America, vol. 99, no. 10, pp. 6883-6888, May. 2002.

[70] N. Nègre et al., "A comprehensive map of insulator elements for the Drosophila genome," *PLoS Genetics*, vol. 6, no. 1, p. e1000814, Jan. 2010.

[71] E. Blackwood and J. T. Kadonaga, "Going the distance: a current view of enhancer action," *Science*, vol. 281, no. 5373, pp. 60-63, Jul. 1998.

[72] M. L. Atchison, "Enhancers: Mechanisms of Action and Cell Specificity," *Annual Review of Cell Biology*, vol. 4, pp. 127-153, 1988.

[73] J. M. G. Vilar and L. Saiz, "DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise," *Current Opinion in Genetics & Development*, vol. 15, no. 2, pp. 136-144, Apr. 2005.

[74] The FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group), "The transcriptional landscape of the mammalian genome," *Science*, vol. 309, no. 5740, pp. 1559-1563, Sep. 2005.

[75] P. Carninci et al., "Genome-wide analysis of mammalian promoter architecture and evolution," *Nature Genetics*, vol. 38, no. 6, pp. 626-635, 2006.

[76] R. V. Davuluri, Y. Suzuki, S. Sugano, C. Plass, and T. H.-M. Huang, "The functional consequences of alternative promoter use in mammalian genomes," *Trends in Genetics*, vol. 24, no. 4, pp. 167-177, Apr. 2008.

[77] R. Leinonen et al., "The European Nucleotide Archive," *Nucleic Acids Research*, vol. 39, no. Database issue, p. D28-D31, Oct. 2011.

[78] P. Flicek et al., "Ensembl 2011," *Nucleic Acids Research*, vol. 39, no. Database issue, p. D800-D806, Nov. 2010.

[79] P. a Fujita et al., "The UCSC Genome Browser database: update 2011," *Nucleic Acids Research*, vol. 39, no. Database issue, pp. 876-882, Oct. 2010.

[80] K. Youens-Clark et al., “Gramene database in 2010: updates and extensions,” *Nucleic Acids Research*, vol. 39, no. Database issue, p. D1085-D1094, Nov. 2010.

[81] C. D. Schmid, R. Cavin Périer, V. Praz, and P. Bucher, “EPD in its twentieth year: towards complete promoter coverage of selected model organisms,” *Nucleic Acids Research*, vol. 34, no. Database issue, p. D82-D85, 2006.

[82] C. Dieterich et al., “Comparative promoter region analysis powered by CORG,” *BMC Genomics*, vol. 6, p. 24, 2005.

[83] N. Sierro, T. Kusakabe, K.-J. Park, R. Yamashita, K. Kinoshita, and K. Nakai, “DBTGR: a database of tunicate promoters and their regulatory elements,” *Nucleic Acids Research*, vol. 34, no. Database issue, p. D552-D555, Jan. 2006.

[84] R. Yamashita, H. Wakaguri, S. Sugano, Y. Suzuki, and K. Nakai, “DBTSS provides a tissue specific dynamic view of Transcription Start Sites,” *Nucleic Acids Research*, vol. 38, no. Database issue, p. D98-D104, Jan. 2010.

[85] K. Tsuchihara et al., “Massive transcriptional start site analysis of human genes in hypoxia cells,” *Nucleic Acids Research*, vol. 37, no. 7, pp. 2249-2263, Apr. 2009.

[86] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott, “NCBI Reference Sequences: current status, policy and new initiatives,” *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D32-6, Jan. 2009.

[87] T. T. Pohar, H. Sun, and R. V. Davuluri, “HemoPDB: Hematopoiesis Promoter Database, an information resource of transcriptional regulation in blood cell development,” *Nucleic Acids Research*, vol. 32, no. Database issue, p. D86-D90, 2004.

[88] R. Gupta, A. Bhattacharyya, F. J. Agosto-Perez, P. Wickramasinghe, and R. V. Davuluri, “MPromDb update 2010: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-seq experimental data,” *Nucleic Acids Research*, vol. 39, no. Database issue, p. D92-D97, Nov. 2011.

[89] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nature Reviews Genetics*, vol. 10, no. 10, pp. 669-680, Oct. 2009.

[90] T. Barrett et al., "NCBI GEO: archive for functional genomics data sets--10 years on," *Nucleic Acids Research*, vol. 39, no. Database issue, p. D1005-D1010, Nov. 2011.

[91] S. K. Palaniswamy, V. X. Jin, H. Sun, and R. V. Davuluri, "OMGProm: a database of orthologous mammalian gene promoters," *Bioinformatics*, vol. 21, no. 6, pp. 835-836, 2005.

[92] A. S. Halees and Z. Weng, "PromoSer: improvements to the algorithm, visualization and accessibility," *Nucleic Acids Research*, vol. 32, no. Web Server issue, p. W191-W194, Jul. 2004.

[93] X. Chen, J.-min Wu, K. Hornischer, A. Kel, and E. Wingender, "TiProD: the Tissue-specific Promoter Database," *Nucleic Acids Research*, vol. 34, no. Database issue, p. D104-D107, Jan. 2006.

[94] T. R. O'Connor, C. Dyreson, and J. J. Wyrick, "Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences," *Bioinformatics*, vol. 21, no. 24, pp. 4411-4413, 2005.

[95] R. T. Morris, T. R. O'Connor, and J. J. Wyrick, "Osiris: an integrated promoter database for Oryza sativa L," *Bioinformatics*, vol. 24, no. 24, pp. 2915-2917, Dec. 2008.

[96] I. A. Shahmuradov, A. J. Gammerman, J. M. Hancock, P. M. Bramley, and V. V. Solovyev, "PlantProm: a database of plant promoter sequences," *Nucleic Acids Research*, vol. 31, no. 1, pp. 114-117, 2003.

[97] Y. Y. Yamamoto and J. Obokata, "Ppdb: a Plant Promoter Database," *Nucleic Acids Research*, vol. 36, no. Database issue, p. D977-D981, 2008.

[98] E. Portales-Casamar et al., “JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles,” *Nucleic Acids Research*, vol. 38, no. Database issue, p. D105-D110, Jan. 2010.

[99] D. Ghosh, “Object-oriented transcription factors database (ooTFD),” *Nucleic Acids Research*, vol. 28, no. 1, pp. 308-310, 2000.

[100] O. L. Griffith et al., “ORegAnno: an open-access community-driven resource for regulatory annotation,” *Nucleic Acids Research*, vol. 36, no. Database issue, p. D107-D113, Jan. 2008.

[101] V. Matys et al., “TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes,” *Nucleic Acids Research*, vol. 34, no. Database issue, p. D108-D110, Jan. 2006.

[102] N. A. Kolchanov et al., “Transcription Regulatory Regions Database (TRRD): its status in 2002,” *Nucleic Acids Research*, vol. 30, no. 1, pp. 312-317, Jan. 2002.

[103] E. Blanco, D. Farré, M. M. Albà, X. Messeguer, and R. Guigó, “ABS: a database of Annotated regulatory Binding Sites from orthologous promoters,” *Nucleic Acids Research*, vol. 34, no. Database issue, p. D63-D67, 2006.

[104] G. Robertson et al., “cisRED: a database system for genome-scale computational discovery of regulatory elements,” *Nucleic Acids Research*, vol. 34, no. Database issue, p. D68-D73, Jan. 2006.

[105] F. Zhao, Z. Xuan, L. Liu, and M. Q. Zhang, “TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies,” *Nucleic Acids Research*, vol. 33, no. Database issue, p. D103-D107, 2005.

[106] A. Yilmaz, M. K. Mejia-Guerra, K. Kurz, X. Liang, L. Welch, and E. Grotewold, “AGRIS: the Arabidopsis Gene Regulatory Information Server, an update,” *Nucleic Acids Research*, vol. 39, no. Database issue, p. D1118-D1122, Jan. 2011.

[107] L. Bülow, Y. Brill, and R. Hehl, "AthaMap-assisted transcription factor target gene identification in Arabidopsis thaliana," *Database*, vol. 2010, p. baq034, Jan. 2010.

[108] A. Guo et al., "DATF: a database of Arabidopsis transcription factors," *Bioinformatics*, vol. 21, no. 10, pp. 2568-2569, 2005.

[109] K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga, "Plant cis-acting regulatory DNA elements (PLACE) database: 1999," *Nucleic Acids Research*, vol. 27, no. 1, pp. 297-300, 1999.

[110] M. Lescot et al., "PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences," *Nucleic Acids Research*, vol. 30, no. 1, pp. 325-327, 2002.

[111] H. Zhang et al., "PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database," *Nucleic Acids Research*, vol. 39, no. Database issue, p. D1114-D1117, Jan. 2011.

[112] P. Pérez-Rodríguez, D. M. Riaño-Pachón, L. G. G. Corrêa, S. A. Rensing, B. Kersten, and B. Mueller-Roeber, "PlnTFDB: updated content and new features of the plant transcription factor database," *Nucleic Acids Research*, vol. 38, no. Database issue, p. D822-D827, Jan. 2010.

[113] W. H. Day and F. R. McMorris, "Critical comparison of consensus methods for molecular sequences," *Nucleic Acids Research*, vol. 20, no. 5, pp. 1093-1099, Mar. 1992.

[114] G. D. Stormo, "DNA binding sites: representation and discovery," *Bioinformatics*, vol. 16, no. 1, pp. 16-23, 2000.

[115] J. E. Stajich et al., "The Bioperl toolkit: Perl modules for the life sciences," *Genome Research*, vol. 12, no. 10, pp. 1611-1618, Oct. 2002.

[116] P. Rice, I. Longden, and A. Bleasby, “EMBOSS: the European molecular biology open software suite,” *Trends in Genetics*, vol. 16, no. 6, pp. 276–277, 2000.

[117] G. Z. Hertz and G. D. Stormo, “Identifying DNA and protein patterns with statistically significant alignments of multiple sequences,” *Bioinformatics*, vol. 15, no. 7-8, pp. 563-577, 1999.

[118] R. Siddharthan, “Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix,” *PLoS ONE*, vol. 5, no. 3, p. e9722, Jan. 2010.

[119] B. Lenhard and W. W. Wasserman, “TFBS: Computational framework for transcription factor binding site analysis,” *Bioinformatics*, vol. 18, no. 8, pp. 1135-1136, Aug. 2002.

[120] G. Thijs et al., “A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling,” *Bioinformatics*, vol. 17, no. 12, pp. 1113-1122, 2001.

[121] J. S. Liu, A. F. Neuwald, and C. E. Lawrence, “Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies,” *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1156-1170, Dec. 1995.

[122] M. K. Das and H.-K. Dai, “A survey of DNA motif finding algorithms,” *BMC Bioinformatics*, vol. 8, no. 7, p. S21, 2007.

[123] T. D. Schneider and R. M. Stephens, “Sequence logos: a new way to display consensus sequences,” *Nucleic Acids Research*, vol. 18, no. 20, pp. 6097-6100, 1990.

[124] M. Tompa et al., “Assessing computational tools for the discovery of transcription factor binding sites,” *Nature Biotechnology*, vol. 23, no. 1, pp. 137-144, Jan. 2005.

[125] J. van Helden, B. André, and J. Collado-Vides, "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies," *Journal of Molecular Biology*, vol. 281, no. 5, pp. 827-842, Sep. 1998.

[126] S. Sinha and M. Tompa, "A statistical method for finding transcription factor binding sites," in *Proceedings of the Eighth International Conference on Intelligent Systems on Molecular Biology*, 2000, vol. 8, no. 1553-0833, pp. 344–354.

[127] T. L. Bailey and C. Elkan, "Unsupervised learning of multiple motifs in biopolymers using expectation maximization," *Machine Learning Journal*, vol. 21, no. 1-2, pp. 51-83, 1995.

[128] T. A. Down and T. J. P. Hubbard, "NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence," *Nucleic Acids Research*, vol. 33, no. 5, pp. 1445-1453, Jan. 2005.

[129] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church, "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation," *Nature Biotechnology*, vol. 16, no. 10, pp. 939-945, 1998.

[130] G. Thijs et al., "A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes," *Journal of Computational Biology*, vol. 9, no. 2, pp. 447-464, Jan. 2002.

[131] M. A. Larkin et al., "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947-2948, Nov. 2007.

[132] E. Berezikov, V. Guryev, R. H. A. Plasterk, and E. Cuppen, "CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting," *Genome Research*, vol. 14, no. 1, pp. 170-178, 2004.

[133] T. Wang and G. D. Stormo, “Combining phylogenetic data with co-regulated genes to identify regulatory motifs,” *Bioinformatics*, vol. 19, no. 18, pp. 2369-2380, Dec. 2003.

[134] R Development Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2011.

[135] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403-410, Oct. 1990.

[136] A. R. Subramanian, M. Kaufmann, and B. Morgenstern, “DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment,” *Algorithms for Molecular Biology*, vol. 3, p. 6, Jan. 2008.

[137] G. Benson, “Tandem repeats finder: a program to analyze DNA sequences,” *Nucleic Acids Research*, vol. 27, no. 2, pp. 573-580, Jan. 1999.

[138] K. L. Adams and J. F. Wendel, “Polyploidy and genome evolution in plants,” *Current Opinion in Plant Biology*, vol. 8, no. 2, pp. 135-141, 2005.

[139] T. Nagy, “Motívum keresés a humán promóterekben,” Pécsi Tudományegyetem, 2011.

[140] C. I. Castillo-Davis and D. L. Hartl, “GeneMerge--post-genomic analysis, data mining, and hypothesis testing,” *Bioinformatics*, vol. 19, no. 7, pp. 891-892, 2003.

[141] D. J. States, W. Gish, and S. F. Altschul, “Improved Sensitivity of Nucleic Acid Database Searches Using Application-Specific Scoring Matrices,” *Methods*, vol. 3, no. 1, pp. 66-70, 1991.

[142] T. Blomme, K. Vandepoele, S. De Bodt, C. Simillion, S. Maere, and Y. Van De Peer, “The gain and loss of genes during 600 million years of vertebrate evolution,” *Genome Biology*, vol. 7, no. 5, p. R43, 2006.

[143] E. Pennisi, “Green Genomes,” *Science*, vol. 332, no. 6036, pp. 1372-1374, 2011.

[144] X. Xie et al., “Systematic discovery of regulatory motifs in human promoters and 3’ UTRs by comparison of several mammals,” *Nature*, vol. 434, no. 7031, pp. 338-345, 2005.

[145] S. Prabhakar et al., “Close sequence comparisons are sufficient to identify human cis-regulatory elements,” *Genome Research*, vol. 16, no. 7, pp. 855-863, 2006.

[146] Q.-fei Wang, S. Prabhakar, S. Chanan, J.-F. Cheng, E. M. Rubin, and D. Boffelli, “Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons,” *Genome Biology*, vol. 8, no. 1, p. R1, 2007.

[147] Y. Li et al., “Genome-wide identification of osmotic stress response gene in Arabidopsis thaliana,” *Genomics*, vol. 92, no. 6, pp. 488-493, Dec. 2008.

[148] N. M. Creux, M. Ranik, D. K. Berger, and A. A. Myburg, “Comparative analysis of orthologous cellulose synthase promoters from Arabidopsis, Populus and Eucalyptus: evidence of conserved regulatory elements in angiosperms,” *The New Phytologist*, vol. 179, no. 3, pp. 722-737, Jan. 2008.

[149] V. Compagnon et al., “CYP86B1 is required for very long chain omega-hydroxyacid and alpha, omega -dicarboxylic acid synthesis in root and seed suberin polyester,” *Plant Physiology*, vol. 150, no. 4, pp. 1831-1843, Aug. 2009.

10. Köszönetnyilvánítás

Szeretnék köszönetet mondani Dr. Barta Endrének, témavezetőmnek, az MBK bioinformatika csoport vezetőjének, akinél elkezdtem bioinformatikus kutatói karrieremet, és aki felhívta a figyelmemet a transzkripciós szabályozás bioinformatikai vizsgálatának fontosságára és kihívásaira.

Köszönetet mondanék Dr. Tóth Gábornak, aki az MBK bioinformatika csoportjának tagjaként sokat segített bioinformatikai tudásom gyarapításában, a Perl programozás elsajátításában. Köszönöm Pálfy Tamásnak és Nagy Tibornak, a MBK bioinformatika csoport további két tagjának a DoOP adatbázis elkészítésében és elemzésében nyújtott segítséget.

Köszönöm Dr. Orosz Lászlónak, hogy annak idején, a Genetikai analízis II című gyakorlat kapcsán a gödöllői bioinformatika csoporthoz irányított, ami alapvetően meghatározta kutatói pályámat.

Végül köszönetet mondanék az MTA-MGKI Alkalmazott Genomika osztályának, ahol disszertációmat és az azzal kapcsolatos munkát befejezhettem, miután elkerültem az MBK-ból.

Rövidítések jegyzéke

API – Alkalmazás fejlesztői felület (*Application Programming Interface*)

BLAST – Lokális illesztéseken alapuló szekvenciahasonlóságot kereső program (*Basic Local Alignment Search Tool*)

BREd – 3’ irányban elhelyezkedő TFIIB kötőhely (*TFIIB Recognition Element downstream*)

BREu – 5’ irányban elhelyezkedő TFIIB kötőhely (*TFIIB Recognition Element upstream*)

CDS – Kódoló szekvencia (*coding sequence*)

ChiP-Seq – Kromatin immunoprecipitációt és szekvenálást kombináló módszer (*Chromatin immunoprecipitation-Sequencing*)

CPAN – Perl modulgyűjtemény (*Comprehensive Perl Archive Network*)

DCE – 5’ irányban elhelyezkedő alap promóter elem (*Downstream Core Element*)

DPE – 3’ irányban található alap promóter elem (*Downstream Promoter Element*)

DRE – DNS replikációhoz kapcsolt alap promóter elem (*DNA Replication-related Element*)

DWM – Dinukleotid súlymátrix (*Dinucleotide Weight Matrix*)

EMBOSS – Molekuláris biológiai szoftvercsomag (*European Molecular Biology Open Software Suite*)

ENA – Európai nukleotid archívum (*European Nucleotide Archive*)

ENCODE – DNS elemek enciklopédiája projekt (*Encyclopedia of DNA Elements*)

ENSEMBL – Genom annotációs adatbázis

HMM – Rejtett Markov modell (*Hidden Markov Model*)

INR – Iniciátor alap promóter elem (*Initiator Element*)

MTE – Motívum 10 alap promóter elem (*Motif Ten Element*)

NCBI – Nemzeti Biotechnológiai Információs Központ, Bethesda, Maryland, USA (*National Center for Biotechnology Information*)

PIC – Transzkripciós preiniciációs komplex (*Preinitiation Complex*)

PSE – snRNS távoli promóter elem (*Proximal Sequence Element*)

PSM – Pozícióspecifikus mátrix (*Position Specific Matrix*)

PSSM – Pozícióspecifikus pontozó mátrix (*Position Specific Scoring Matrix*)

PWM – Pozícióspecifikus súlymátrix (*Position Weight Matrix*)

RefSeq – NCBI referencia szekvencia adatbázis (*Reference Sequence*)

snRNS – kis nukleáris RNS (*small nuclear RNA*)

TBP – TATA-box kötő fehérje (*TATA-binding protein*)

TF – Transzkripciós faktor (*Transcription Factor*)

TFBS – Transzkripciós faktor kötőhely (*Transcription Factor Binding Site*)

TRF – Tandem ismétlődéseket kereső szoftver (*Tandem Repeats Finder*)

TSS – Transzkripciós starthely (*Transcription Start Site*)

UTR – Nem transzlálódó régió (*Untranslated Region*)

XCPE1 – Hepatitis B X gén alap promóter elem (*X gene Core Promoter Element 1*)

Ábrák jegyzéke

Táblázatok jegyzéke

Online mellékletek jegyzéke

Az online mellékletek a http://doop.abc.hu/download/ címen érhetők el.

01-plant-vALL-cluster-number.csv A növényi adatbázis verziók ortológ csoportjainak száma.

02-plant-vALL-sequence-number.csv A növényi adatbázis verziók szekvenciáinak száma.

03-plant-vALL-sequence-number-per-cluster.csv A növényi adatbázis verziók ortológ csoportjainak szekvenciaszáma.

04-chordate-vALL-cluster-number.csv A gerinces adatbázis verziók ortológ csoportjainak száma.

05-chordate-vALL-sequence-number.csv A gerinces adatbázis verziók szekvenciáinak száma.

06-chordate-vALL-sequence-number-per-cluster.csv A gerinces adatbázisverziók ortológ csoportjainak szekvenciaszáma.

07-plant-v18-subset-number.csv Az 1.8-as növényi adatbázis ortológ alcsoportjainak száma.

08-plant-v18-sequence-number-per-subset.csv Az 1.8-as növényi adatbázis ortológ alcsoportjainak szekvenciaszáma.

09-plant-v18-taxons-per-subset.csv Az 1.8-as növényi adatbázis fontosabb fajainak száma.

10-plant-v18-subset-with-motif-percent.csv Az 1.8-as növényi adatbázis motívumot tartalmazó alcsoportjainak százalékos értékei.

11-plant-v18-motif-number.csv Az 1.8-as növényi adatbázis ortológ alcsoportjainak összesített motívumszáma.

12-plant-v18-motif-number-per-subset.csv Az 1.8-as növényi adatbázis ortológ alcsoportjainak motívumszáma.

13-plant-v18-motif-size-total.csv Az 1.8-as növényi adatbázis ortológ alcsoportjainak összesített motívumhossza.

14-plant-v18-motif-length-uniq-per-subset.csv Az 1.8-as növényi adatbázis ortológ alcsoportjainak motívumhossza.

chordate-v1.0-data.tar.gz A gerinces adatbázis 1.0 verziójának szekvenciái és motívumai.

chordate-v1.1-data.tar.gz A gerinces adatbázis 1.1 verziójának szekvenciái és motívumai.

chordate-v1.3-data.tar.gz A gerinces adatbázis 1.3 verziójának szekvenciái és motívumai.

chordate-v1.4-data.tar.gz A gerinces adatbázis 1.4 verziójának szekvenciái és motívumai.

plant-v1.0-data.tar.gz A növényi adatbázis 1.0 verziójának szekvenciái és motívumai.

plant-v1.2-data.tar.gz A növényi adatbázis 1.2 verziójának szekvenciái és motívumai.

plant-v1.3-data.tar.gz A növényi adatbázis 1.3 verziójának szekvenciái és motívumai.

plant-v1.5-data.tar.gz A növényi adatbázis 1.5 verziójának szekvenciái és motívumai.

plant-v1.8-data.tar.gz A növényi adatbázis 1.8 verziójának szekvenciái és motívumai.

mofext-1.0.3.tar.gz A MOFEXT program forráskódja.

Bio-DOOP-DOOP-1.04.tar.gz A DOOP API forráskódja.

Az értekezéshez kapcsolódó közlemények jegyzéke

Referált tudományos folyóiratokban megjelent cikkek

E. Sebestyén, T. Nagy, S. Suhai, and E. Barta, „DoOPSearch: a web-based tool for finding and analysing common conserved motifs in the promoter regions of different chordate and plant genes," *BMC Bioinformatics*, vol. 10, no. S6, Jun. 2009

E. Barta, E. Sebestyén, T.B. Pálfy, G. Tóth, C.P. Ortutay, and L. Patthy, „DoOP: Databases of Orthologous Promoters, collections of clusters orthologous upstream sequences from chordates and plants," *Nucleic Acids Research*, vol. 33, no. Database issue, pp. D86-D90., Jan. 2005

Egyéb cikkek

E. Sebestyén, T. Nagy, T. Pálfy, G. Tóth, and E. Barta, „Comparative genomics approach in promoter analysis," *EMBNet News*, vol. 12, no. 2, pp. 23-26., Dec. 2006

Poszter összefoglalók konferencia kiadványokban

E. Sebestyén and E. Barta, „Identifying conserved promoter motifs and transcription factor binding sites in orthologous plant promoter collections," *Plant Breeding and Biotechnology in the Great Pannonian Region*, Cluj-Napoca, Romania, Jul. 4-7 2010

E. Sebestyén, T. Nagy, T. Pálfy, G. Tóth, and E. Barta, „Identifying common conserved promoter motifs between genes, using the taxonomic group-based motif collections of the DoOP database," *15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB)*, Vienna, Austria, Jul. 21-25 2007

Sebestyén E., Nagy T., Pálfy T., Szenes Á., Molnár J., Tóth G., és Barta E., „A transzkripciós szabályozás vizsgálata bioinformatikai módszerekkel – A DoOP adatbázis és a DoOPSearch keresőoldal," *Magyar Biokémiai Egyesület 2006. évi vándorgyűlése*, Pécs, 2006 aug. 30. - szept. 2.