TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Arvutiteaduse instituut
Informaatika õppekava

Andreas Ellervee

# Geeniregulatsiooni signaali vahendavate geenide automaatne leidmine suuremahulistest eksperimentaalsetest andmetest

Bakalaureuse töö (9 EAP)

Juhendaja: Hedi Peterson, MSc
Juhendaja: Elena Sügis, MSc

Tartu 2015

# Geeniregulatsiooni signaali vahendavate geenide automaatne leidmine suuremahulistest eksperimentaalsetest andmetest

**Lühikokkuvõte**

Käesoleva bakalaureusetöö eesmärgiks on automaatselt leida geenidevahelisi seoseid suuremahulistest eksperimentaalsetest andmetest. Olenevalt eksperimendi tüübist saame me tihtipeale teada vaid osa tegelikest regulatoorsetest seostest geenide vahel. Kombineerides erinevaid eksperimente saame aga täpsemini määratleda otseseid ja kaudseid seoseid, mis määravad ära rakkude seisundi mingil kindlal tingimusel.

Geenide häirituse (ingl k. *perturbation*) eksperimendiga on võimalik leida need geenid, mida häiritud geen mõjutab, kas siis otseselt või kaudselt. Kui neid andmeid kombineerida kromatiinisadestamise andmetega, siis saame eraldada otseselt mõjutatud geenid kaudselt mõjutatutest. Käesoleva töö eesmärk ongi leida kaudselt mõjutatud geenidele võimalikke regulaatoreid, mis vahendaksid signaali häiritud geenilt.

Signaalivahendajate leidmiseks kombineerime ENCODE andmestikus olevaid andmeid geenide võimalike regulaatorite ja aktiivsete promootorite kohta. Juhul kui geeni regulatoorses alas on näidatud transkriptsioonifaktori seondumist, siis peame seda võimalikuks signaalivahendajaks peamiselt regulaatorilt. Edasi tuleb leida need potentsiaalsed signaalivahendajad, mis on ise peamise regulaatori poolt otseselt mõjutatud. Töö tulemusena tekivad kolmikud geenidest, kus peamine regulaator reguleerib otseselt signaalivahendajat. See omakorda reguleerib peamise regulaatori sihtmärkgeene teatud bioloogilistel tingimustel.

Töö lõppeesmärgiks on kasutatav tööriist, mis võimaldab sobivate sisendandmete ja bioloogiliste tingimuste määramisel automaatselt sellised regulatoorseid kolmikuid leida.

**Märksõnad:** geen, transkriptsioonifaktor, geeniregulatsioon, ENCODE, geeniregulatsiooni signaalivahendajad

# Systematic approach for finding gene regulation signal mediators from large-scale experimental data

**Abstract**

The goal of the thesis is to automatically find relations between genes from large heterogeneous experimental data. Depending on the type of the experiment, we can only find out very little about the actual regulation between genes. By combining different experiments we can determine the direct and indirect regulation of genes much more accurately.

With gene perturbation experiments we can identify genes, that the perturbed gene affects either directly or indirectly. If this data is combined with chromatin immunoprecipitation data, then we can separate directly affected genes from indirectly affected ones. The aim of this thesis is to find the intermediate regulators that carry the signal from the perturbed gene to the indirectly affected genes.

To find these intermediate regulators we combine ENCODE data with possible regulators and active promoters of genes. In case the gene's regulatory area is shown to bind with a transcription factor, then we consider it a possible signal transmitter from the main regulator. As the end result, we will have regulatory triplets, where main regulator regulates an intermediate signal transmitter, that in return regulates the target gene.

The primary target of this thesis is to build a tool, which can find these regulatory triplets automatically given the proper input data.

**Keywords:** gene, transcription factor, gene regulation, ENCODE, intermediate regulators

# Sisukord

# Sissejuhatus

Üheks elu põhiomaduseks on pärilikkus, mis avaldub kõikides bioloogilistes funktsioonides ja struktuurides. Kõige paremini avaldub pärilikkus organismide paljunemise protsessis. Pärilikkuse sisu on geneetiline informatsioon ning see kujutab endast teavet valkude ehituse ja struktuuri, kuid samuti ka nende kasutamise kohta. Geneetiline info peitub aga geenides.

Aastal 2003 määras Human Genome Project (eesti k. *Inimese Genoomi Projekt*) täielikult esimese inimese genoomi ning leiti, et inimeses on ligikaudu 20000 geeni. Kuigi geenide kaardistamine õnnestus, ei ole kaugeltki veel teada kõikide geenide funktsionaalsus ning just see ongi järgmine suur samm inimese genoomi uurimises - koostada geenide regulatoorsed võrgustikud ning koguda informatsiooni geeni funktsioonide kohta. Selleks on vaja täpsemalt teada geenide ja valkude seoseid. Meile huvipakkuvateks valkudeks on transkriptsioonifaktorid, mis seonduvad kindlatele DNA järjestustele, mida nimetatakse promootoriks, ning reguleerivad mRNA (ingl. k. *Messenger-RNA*) transkriptsiooni taset.

Transkriptsioonifaktorite (TF) seondumiste kohta saame informatsiooni kromatiini immunosadestamise (ChIP-*) katsega. TF-id saavad seonduda ainult siis, kui DNA ei ole seotud ümber histoonide (valgud, mis seovad DNA enda ümber, et seda kompaktsemalt kromosoomis hoiustada) ehk on avatud - seda on katseliselt võimalik mõõta DNase I katsetega.

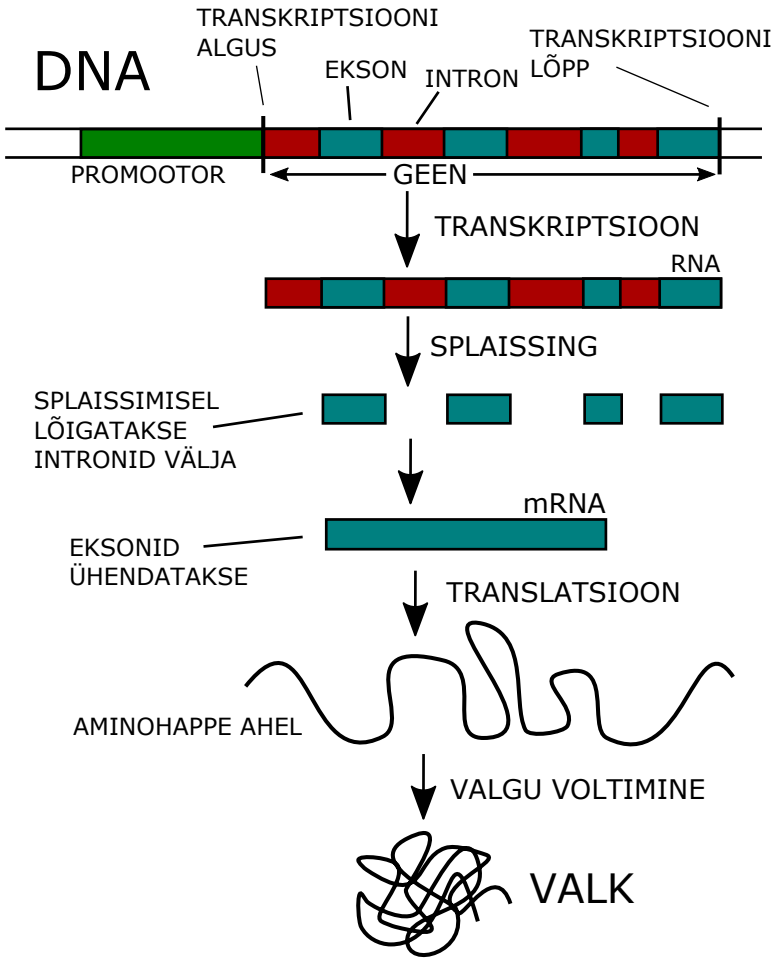
Transkriptsioonifaktori seondumine DNA-ga on ainult üks osa geeniregulatsioonist. Tihtipeale TF-id seonduvad, aga ei algata transkriptsiooni, sest neil on vaja mingit aktivaatorit. Seega uuritakse geenide häirituse katsega, millised on täpsemalt need geenid (valgud), millel on oluline roll regulatsioonis teatud tingimustel. Häirituse katses mõjutatakse ühte geeni kas positiivselt või negatiivselt ja mõõdetakse, millised on teised geenid ja kui suurel määral muutub nende ekspressiooni tase.

Kombineerides geenide häirituse eksperimendist saadud andmeid ChIP-* katsest saadud DNA-ga seondumise andmetega, on võimalik eristada kaht gruppi sihtmärkgeene - neid, mida mõjutatakse läbi DNA-ga seondumise ning neid, mida mõjutavad mingid vahefaktorid ehk signaalivahendajad. Antud töö eesmärgiks ongi uurida juhtumeid, kus signaal liigub mõjutatud geeni sihtmärkidelt teistele geenidele ning leida need signaalivahendajad.

# 1 Bioloogiline taust

Inimese kehas on ligikaudu 50 miljardit rakku. Inimene on päristuumse rakutüübiga organism, mis tähendab, et pärilik informatsioon asub rakutuumas kromosoomidena. Inimesel on 23 kromosoomipaari ning nendesse on jaotunud üks koopia DNA-st. Igas kromosoomis on lineaarses järjestuses kindlalt paiknevad geenid. Geenid on funktsionaalsed lõigud DNA-s, mis hoiavad endas infot organismi ülesehituse kohta. Seda informatsiooni kasutatakse valkude sünteesiks ning valgud ongi inimese keha nii-nimetatud ehitusmaterjal. Igal organismil on rohkesti geene, mis vastavad erinevatele bioloogilistele tunnustele nagu näiteks juuste või silmade värv, aga ka tunnustele mis pole silmaga nähtavad, nagu näiteks veregrupp. Kogu see geenides peituv informatsioon realiseeritakse geeniekspressioonil, millest antud peatükis juttu tuleb. Samuti kirjeldatakse geeni transkriptsiooni ning kuidas see töötab ja miks see elusorganismis oluline on.

## 1.1 Geenieksressioon



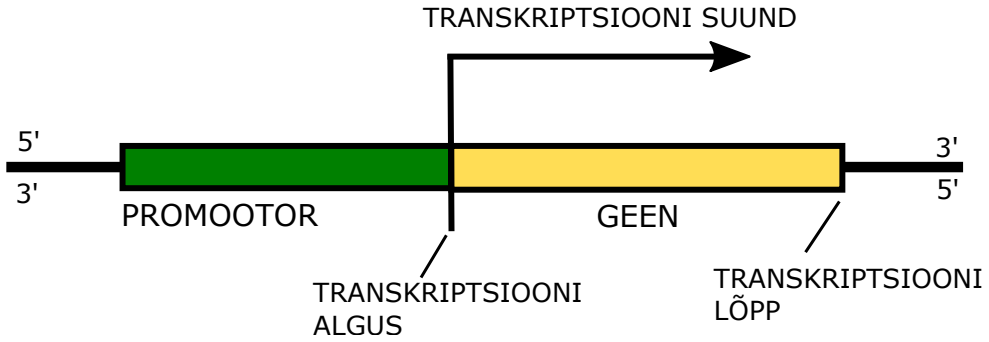

**Joonis 1:** Geeni ekspressiooni etapid. Transkriptsioon, splaissing ja translatsioon. Transkriptsiooni käigus tehakse DNA ahelale vastav RNA ahel. Splassinguga lõigatakse RNA-st välja intronid ning saadakse mRNA. mRNA transleeritakse aminohappe ahelaks mis omakorda voldib end valguks. Valk on geeni lõpp-produkt.

Geenieksрessioon (joonis 1) on elusorganismi geneetilise informatsiooni avaldumine: protsess, mille käigus geenides sisalduv pärilik materjal avaldub geeni lõpp-produktina. Nendeks produktideks on enamasti valgud ehk proteiinid. Valgud moodustavad inimkehast ligi 20%. Valke ei kasutata peamise energiaallikana, vaid nad on naha luude ja teiste kudede alustalaks. Seega kasutatakse neid pigem organismi ehitamiseks, kudede taastamiseks ja kahjustunud rakkude uuendamiseks.

Geenieksрessioonil on mitu sammu (joonis 1). Esimene etapp on geeni DNA lõigule vastava RNA ahela kodeerimine. DNA koosneb nukleotiitsetest aluspaaridest G - C ja A - T (vastavalt guaniin - tsütosiin ja adeniin - tümiin). RNA sünteesimise käigus transkribeeritakse vastav RNA ahel, kus T (tümiin) on asendunud U-ga (uratsiiliga) [4]. Nii geenid kui ka kodeeritud RNA ahel sisaldavad eksoneid ja introneid. Ekson on valku kodeeriv järjestus genoomis, samas kui intronite järjestust valkude sünteesimiseks ei kasutata. Intronid lõigatakse geenieksрessiooni teises etapis, splaissimisel, välja. Tänu alternatiivsele splaissimisele, kus ühe geeni erinevate eksonite komplekt toodab erinevaid valke, on võimalik ligikaudu 20 000 geenist toota rohkem kui 100 000 valku. Eksonid ühendatakse ning tulemuseks on mRNA ehk informatsiooni-RNA, mis järgmisena transleeritakse aminohappe ahelaks. mRNA-s paiknevad nukleotiidid kolmikutena, kus iga kolmik sümboliseerib üht kindlat aminohapet [1]. Aminohappe ahel on valgu esmane struktuur. Pärast translatsiooni toimub modifikatsioon, kus aminohappe ahel võtab kolmedimensioonilise kuju ning tulemuseks ongi funktsionaalne valk. Antud töö juures huvitab meid kõige rohkem geenieksрessiooni esimene etapp ehk transkriptsioon.

## 1.2 Transkriptsioon

Transkriptsioon on geeni avaldumise esimene etapp, kus teatud osa DNAst transkribeeritakse ümber RNA lõiguks, mida kutsutakse ka transkriptsiooniühikuks.




**Joonis 2:** Geeni piirkond eukarüoodis ehk päristuumses organismis. Rohelisega on tähistatud promootori ala, kuhu seonduvad transkriptsioonifaktorid ja mRNA polümeraas. Kollasega on tähistatud geen, mida mRNA polümeraas hakkab transkribeerima ehk looma DNA-le vastavat RNA ahelat.

Geene transkribeeritakse raku tuumas, kus paikneb ka DNA. Igas rakus on sama koopia DNAst, kuid selle transkriptoom (ingl. k. *Transcriptome*) ehk kõikvõimalikud transkribeeritud RNA-d, erineb sõltuvalt funktsioonist ja rakutüübist. Näiteks insuliini tootvad rakud pankreases (kõhunäärmes) sisaldavad transkripte insuliini jaoks, mida seevastu luude rakud ei sisalda [3]. Kuigi luude rakud kannavad sama geeni, siis seda ei

transkribeerita, sest et nii-öelda bioloogilised tingimused pole sobilikud. Lihtsamalt öeldes, transkriptoom kirjeldab, millised geenid avalduvad antud rakus teatud tingimustel.

Geenide transkriptsioon toimub kolmes etapis ning transkriptsioonifaktori olemasolu on väga oluline [4]. Transkriptsiooni esimeses etapis kinnitub ensüüm nimega RNA polümeraas kindla DNA järjestuse külge, mida nimetatakse promootoriks (joonis 2). Tihti transkriptsioonifaktorid seonduvad, aga ei alusta teatud tingimusel transkriptsiooni. Sellisel juhul on süsteemis kaudne regulaator, mis aktiveerib antud transkriptsiooni - sellest täpsemalt kirjas järgmises alapeatükis.

Teiseks transkriptsiooni etapiks nimetatakse elongatsiooni. DNA kuju on küll kahe ahelaga (topelt-heeliks), aga antud etapis läheb vaja ainult ühte ahelat, mida nimetatakse RNA sünteesi malliks (kollane piirkond joonisel 2). Transkriptsioon jätkub ning RNA polümeraas hakkab kodeerima ehk sünteesima uut RNA ahelat. RNA polümeraas liigub mööda DNA-d suunaga $5' \rightarrow 3'$ (joonisel 2 on noolega välja toodud transkriptsiooni suund). Loodud RNA ahelas on T (tümiin) asendunud U-ga (uratsiil).

Kolmandaks ehk viimaseks transkriptsiooni etapiks on terminatsioon. Et uue RNA ahela süntees ei kestaks igavesti, on vajalik terminatsiooni signaal, mis transkriptsiooni kompleksi lagundaks. Kui värskelt sünteesitud RNA molekul moodustab G-C rikka struktuuri, millele järgnevad nõrgad U-A sidemed, siis mehaanilise surve tõttu nad lõhutakse ning RNA polümeraas tõmmatakse transkripti aktiivpiirkonnast välja. See põhjustab transkriptsiooni terminatsiooni.

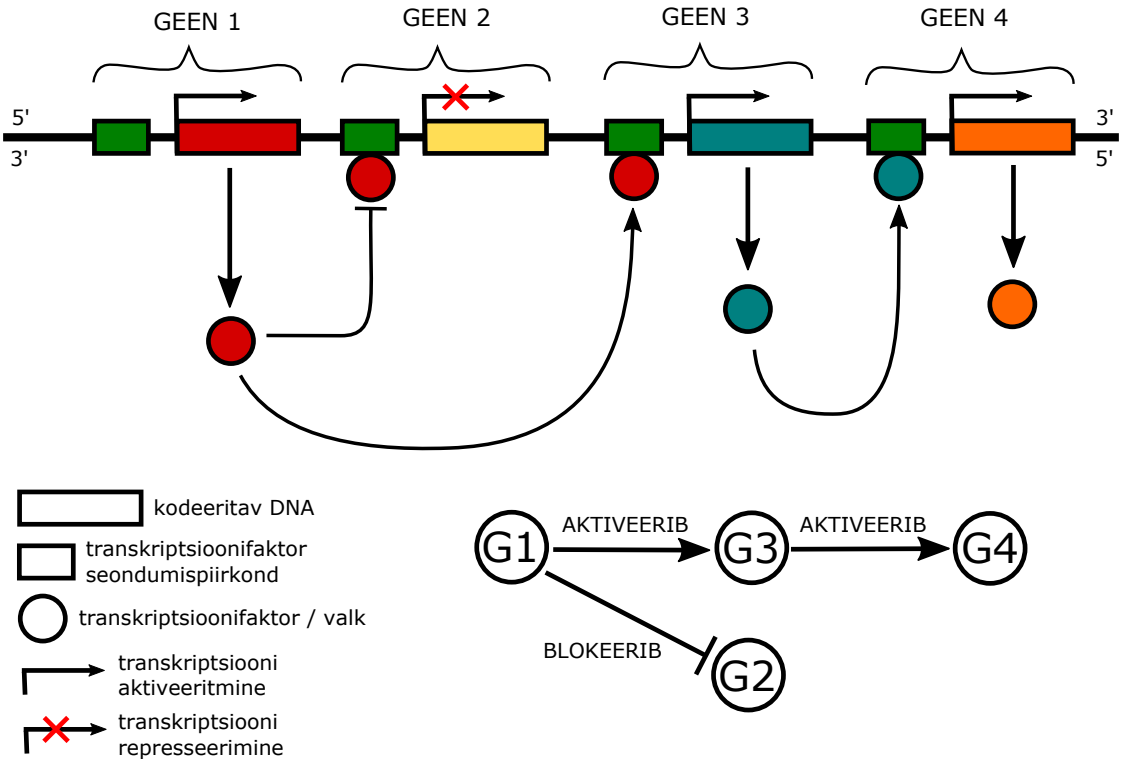
### 1.2.1 Transkriptsioon inimeses

Suurte organismide puhul, nagu näiteks inimesed, ei ole geenide regulatsioon kaugeltki nii selge ning tegemist on palju keerulisema süsteemiga. Esiteks, inimese DNAs ei paikne geenid järjest ning nende vahel on palju "rämps-DNAd"ehk osa DNAst, millel puudub teadaolev funktsionaalsus. RNA polümeraas ei saa vabalt seonduda promootori piirkonnaga, selleks on vaja transkriptsioonifaktorite olemasolu, mis aitavad RNA polümeraasil seonduda promootori piirkonnaga.

Lisaks ei ole teada, millist geeni täpsemalt mingisugune seondumispiirkond reguleerib. Kui DNA lõigus on palju geene koos, on potentsiaalseid sihtmärke väga palju [8]. Enamik regulatoorsetest valkudest mõjutavad mitme geeni transkriptsiooni. See juhtub ka siis, kui antud regulatoorsel valgul eksisteerib mitu seondumispiirkonda raku genoomis. Järelikult võib reguleeriv valk mängida rolli erinevate geenide ekspressioonis [3].

## 1.3 Geenide regulatsioon

Kuna osad regulatoorsed valgud seonduvad mitmesse eri piirkonda genoomil, siis järelikult võivad nad ka mõjutada mitme geeni ekspressiooni. Samuti ühe geeni ekspressioon võib mõjutada või blokeerida teise geeni avaldumist.




**Joonis 3:** Geenide regulatsiooni mudel. Rohelisega on tähistatud transkriptsioonifaktori seondumispiirkond. On näha, kuidas GEEN 1 produktiks on regulatoorne valk, mis blokeerib GEEN 2 ekspressiooni, aga aktiveerib GEEN 3, mis omakorda reguleerib GEEN 4. Antud töö eesmärgiks on leida kõik GEEN 3 sarnased signaalivahendajad.

Joonisel 3 on näidatud geeni regulatsiooni võrgustiku mudel. See mudel ei ole täpne, kuid annab ettekujutuse, kuidas toimub geenide niinimetatud 'multifaktoriaalne regulatsioon' ehk regulatsioon, kus osaleb rohkem kui üks geen. Jooniselt on näha, et geen nr. 1 produktiks on valk, mis seob end geen nr. 2 ja geen nr. 3 promootori piirkonda. Järelikult üks geen mängib rolli kahe teise geeni ekspressioonis. Nimelt geen nr. 1 blokeerib geeni nr. 2 transkriptsiooni ja reguleerib 3. geeni, mis omakorda avaldab mõju 4. geenile. Olgugi, et katsete tulemusena saame teada, et geeni nr. 1 mõjutamisel on näha muutust geen nr. 4 ekspressioonis, ei saa me täpselt teada, mis geenid on selle regulatsiooni vahelülideks. Erinevate katsetulemuste kombineerimisel on aga võimalik luua hüpoteese, millised geenid võivad olla regulatsioonis signaalivahendajateks ning hiljem seda katseliselt kinnitada või kummutada.

## 1.4 Rakuliin

Rakuliin on rakkude kogum organismidelt, mis tavapäraselt ei paljuneks, kuid tänu rakumutatsioonile ja tehislikule keskkonnale on nad eemaldanud normaalsest raku vananemise protsessist ning suudavad jätkata paljunemist *in vitro*, mis tagab rakkude edasise kasvu. Sellist säilitamist kasutatakse põhiliselt katselistel eesmärkidel.

Rakuliine saab teoreetiliselt lõpmatuseni kloonida, mis annab eelise teha analüüse ja katseid geneetiliselt samadel rakkudel. See on oluline korduvates teaduslikes eksperimentides, et testida näiteks eukarüootide geenide ekspressiooni erinevatel tingimustel.

Kõige esimene ja siiani kasutuses olev rakuliin on HeLa [13], mis saadi emakakaelavähist. Embrüonaalsete tüvirakkude uurimisel kasutatakse peamiselt rakuliine H1 ja H9. ENCODE andmestikus on kokku tehtud katseid 91 erineva rakuliiniga [2].

## 1.5 Transkriptsioonieksperimendid

Tänu genoomi järjestamisele on teada, kus paiknevad geenid. Samuti on erinevatest eksperimentaalsetest allikatest teada milliste promootorpiirkondadega erinevad transkriptsioonifaktorid seonduvad. Tihtipeale on need andmed aga üldised ning kindla bioloogilise tingimuse kirjeldamiseks on vaja täpsemalt välja selgitada aktiivsete geenide regulatsioonimehhanismid.

### 1.5.1 DNase I katse

Kromatiin on DNA ja sellega seondunud valkude, histoonide, kogum, millest moodustuvad kromosoomid. Kromosoomis on DNA keerunud ümber valkude, mida kutsutakse histoonideks. Selleks, et transkriptsioonifaktor saaks DNA-ga seonduda, on vaja, et DNA oleks antud kromatiinis avatud - see tähendab, et DNA ei oleks histoonide ümber tihedalt seotud, sest muidu ei oleks ruumi transkriptsioonifaktoril seonduda. Sellisel viisil on ka histoonid ja kromatiini olek seotud geeniregulatsiooniga.

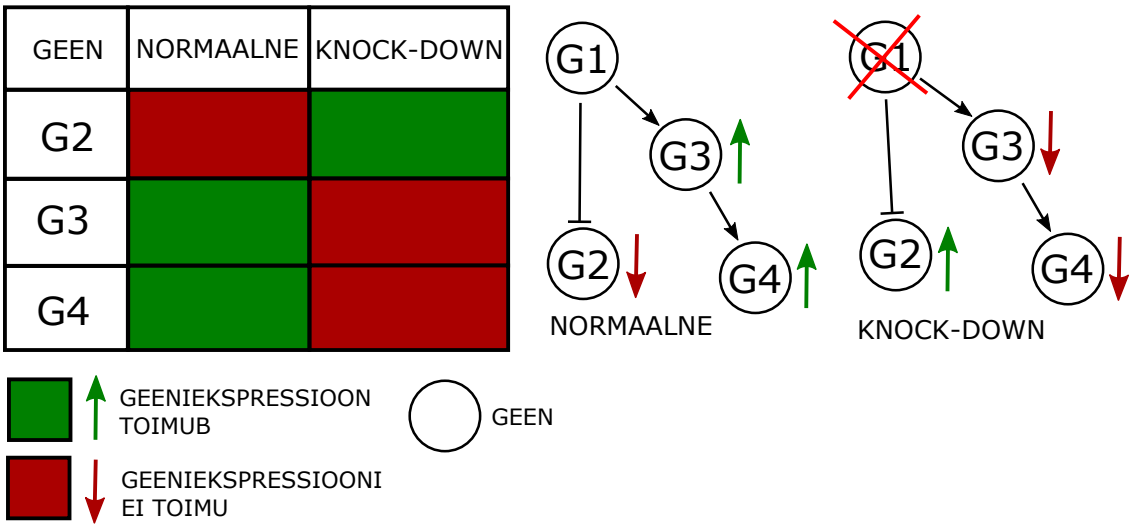
Selliseid piirkondi on võimalik mõõta ensüümi deoksüribonukleaas I, DNase I, katsetega. DNase I lõikab DNA ahelat enamasti valkudega mitte seotud piirkondadest. Sellistesse piirkondadesse on enamasti seondunud transkriptsioonifaktorid. Millised transkriptsioonifaktorid antud avatud piirkondadesse aga seonduvad, on võimalik määrata kromatiini immunosadestamisega.

### 1.5.2 Kromatiini immunosadestamine

Kromatiini immunosadestamine on üks tavapärane viis, kuidas leida huvipakkuva transkriptsioonifaktori seondumispiirkondi DNA-s. See tehnika annab aimdust, millised on geenide ja valkude omavahelised suhted rakus.

Kromatiini immunosadestamine näeb lühidalt välja selline, et huvipakkuv valk ja kromatiin seostatakse ajutiselt. DNA jagatakse ligikaudu 500 aluspaari suurusteks osadeks ultrahelitöötluse abil. Need osad, kus huvipakkuv proteiin asetseb, sadestatakse kasutades sellele proteiinile omast antikeha [9]. Sellega seotud DNA lõigud puhastatakse ja selgitatakse välja nende sekventsid ehk järjestused ja paigutatakse nad selle abil genoomile. Leitud sekventsid ehk DNA lõigud loetakse selle valguga seostunuks.

### 1.5.3 Geenide häirituse eksperiment




**Joonis 4:** Geenide häirituse näide geeniregulatsiooni mudeli põhjal (joonis 3). Roheline värv tähistab geeni avaldumist, punane värv tähendab, et geeni ekspressiooni ei toimu. Tulbas *"NORMAALNE"* on toodud geenide G2, G3 ja G4 normaalsed olekud. Tulbas *"KNOCK-DOWN"* on toodud samade geenide olekud, kus huvipakkuvat geeni G1 on häiritud ehk ekspressiooni taset on vähendatud. Näiteks geen G3 avaldub normaalsetes tingimustes, aga kui huvipakkuv faktor G1 inhibeerida (ingl. k. *inhibit*), siis on näha, et geeni G3 ekspressiooni tase on langenud ja geenieksрessiooni ei toimu. Vastupidiselt tavaolekus G2 ei avaldu, sest G1 blokeerib G2 ekspressiooni. Kui aga häirida meile huvipakkuvat geeni, siis on näha, et G2 ekspressiooni tase tõuseb.

Geenide häirituse eksperiment seisneb selles, et huvipakkuval transkriptsioonifaktoril suurendatakse või vähendatakse ekspressiooni taset (ingl. k. vastavalt *upregulation* ja *downregulation*) nii, et sihtmärkgeenid näitavad märkimisväärset muutust ekspressioonis [12]. Selle abil saame jälgida, millised geenid ei hakka end üldse ekspresseerima või just vastupidi - geeni ekspressiooni tase on märgatavalt tõusnud. Sellist eksperimenti tuntakse kui transkriptsioonifaktori häirituse eksperiment (TFPE - ingl. k. *Transcription Factor Perturbation Experiment*).

# 2 Praktiline töö

Järgnevas peatükis tuleb juttu töö käigust. Milliseid andmeid on kasutatud ning mis on nende andmete päritolu. Samuti kuidas neid omavahel kombineerida, et kaardistada geenid ja nende lähedusümbruses asuvad transkriptsioonifaktorite seondumispiirkonnad.

## 2.1 Bioloogilised katsed

Molekulaarbioloogias tehakse väga palju katseid tranksriptsioonifaktorite ja DNA-ga, täpsemalt uurida inimese bioloogiliste protsesside regulatsiooni. Kuna katseid on palju, siis on ka rohkelt andmeid, mis katsetest tulevad. Kuna katsed on tihtipeale eraldiseisvad, siis ühe katse tulemuste juures ei arvestata teise katsega. Sellised erinevad tulemused ei anna täielikku arusaama bioloogilistest protsessidest.

Mõistmaks paremini erinevaid bioloogilisi protsesse, tuleks katsetulemustest saadud andmeid omavahel integreerida ja vaadata neid kui ühtse tervikuna, et saada täpsem ülevaade peituvast regulatoorsest funktsionaalsustest.

## 2.2 Üldised andmestikud

Katsetest saadud andmestikku võib igat üht vaadata kui eraldi kihti, mida üksteise peale ladudes jõuame uute teadmisteni, mis on vanadest täpsemad ja paremad.

### 2.2.1 GENCODE

ENCODE[1] (ingl. k. *The Encyclopedia of DNA Elements*) on aastal 2003 loodud rahvusvaheline projekt, mille eesmärgiks on olnud kaardistada inimese genoomi kõik funktsionaalsed elemendid. GENCODE[2] (ingl. k. *Encyclopedia of genes and gene variants*) on ENCODE projekti alamprojekt, mille eesmärk on kaardistada kõik valke kodeerivad geenid ja tuvastada geenide funktsionaalsust inimestes ja hiirtes. GENCODE andmestikus sisaldub informatsioon järgneva kohta:

- Kromosoom
- Kirjeldus
- Funktsionaalse elemendi tüüp
- Alguspunkti koordinaat kromosoomil
- Lõpp-punkti koordinaat kromosoomil
- DNA strand, kus "+"tähistab geeni kodeerimist suunal $5' \rightarrow 3'$ ja "−"tähistab kodeerimist suunal $3' \rightarrow 5'$
- Geeni Ensembl ID

[1] https://www.encodeproject.org/
[2] http://www.gencodegenes.org/

### 2.2.2 Geeni nimed

*g:Profiler*[3] [10] on avalik veebirakendus, mida kasutatakse geenide iseloomustamiseks. *g:Profileril* on alam-tööriist *g:Convert*, mida kasutatakse geenide erinevate identifikaatorite ja nimede tõlkimiseks. *g:Convert* põhineb Emsembl-i andmebaasil ning on võimeline tõlkima palju eri liikide nimetüüpe.

- Geeni Ensembl ID
- Geeni nimi

### 2.2.3 DNase I

ENCODE projekti raames on loodud DNase I andmestik[4], mis sisaldab informatsiooni hüperaktiivsetest seondumispiirkondadest. Sellised on piirkonnad, kus DNA ei ole kromosoomis histoonide ümber keerdunud. Sinna saavad transkriptsioonifaktorid seonduda ning mängida rolli geenide ekspressioonis. DNase I andmestik sisaldab endas järgmist infot:

- Kromosoom
- Piirkonna alguskoordinaat kromosoomil
- Piirkonna lõppkoordinaat kromosoomil
- DNA ahel, kus "+"tähistab $5' \rightarrow 3'$ suunda ja "−"tähistab $3' \rightarrow 5'$ suunda
- Kodeerimispiirkonna alguskoordinaat (sama, mis piirkonna alguskoordinaat)
- Kodeerimispiirkonna lõppkoordinaat (sama, mis piirkonna lõppkoordinaat)
- Rakuliinide loend
- Rakuliinide arv

### 2.2.4 ChIP-seq

ENCODE projektis on saadaval veel ka ChIP-seq andmestik, mis on saadud kromatiini immunosadestamise ja DNA sekveneerimise tulemusena. Antud andmestikus on ära kaardistatud transkriptsioonifaktorite seondumispiirkonnad ning täpsustatud, millised faktorid on seal seondunud. ChIP-seq andmestik sisaldab endas järgnevat:

- Kromosoom
- Seondumispiirkonna alguskoordinaat kromosoomil
- Seondumispiirkonna lõppkoordinaat kromosoomil
- Transkriptsioonifaktorite nimede loend (pikema loendi puhul märgitakse *multiple*)
- Seondunud faktorid ja rakuliin kujul *Transkriptsioonifaktor(Rakuliin)*

[3] biit.cs.ut.ee/gprofiler/welcome.cgi
[4] https://www.encodeproject.org/data/annotations/

### 2.2.5 Geenide häirituse katse andmed

Kuna ChIP-seq katse annab aimdust, millised transkriptsioonifaktorid genoomis seonduvad, ei määra see ära, kas geen ekspresseerub või mitte. Geenide häirituse katsest 1.5.3 saame andmestiku, kus on märgitud geenid, mis näitavad märkimisväärset muutust geeniekspressioonis meile huvipakkuva geeni häirimise korral.

Geenide häirituse andmestik sisaldab:

- Geeni Ensembl ID
- Geeni nimi

## 2.3 Huvipakkuva faktori spetsiifilised andmestikud

Regulatoorsete signaalivahendajate leidmiseks tuleb kombineerida huvipakkuva faktori otseselt ja kaudselt mõjutatud geenide andmestikke üldiste andmestikega. Näitena on toodud *OCT4* (tuntud ka kui *POU5F1*) poolt otseselt ja kaudselt mõjutatud geenide andmestikud rakuliinil H1-hESC. OCT4 on üks kolmest embrüonaalsete tüvirakkude peamisest transkriptsioonifaktorist, mis vastutavad rakkude pluripotentsuse ja eneseuuenduse eest. Need andmed on pärit *Embryonic Stem Cell Database*[5] (ESCDb) [7] andmebaasist. ESCDb baasis on inimese ja hiire embrüonaalsete tüvirakkudega seotud andmestikke, samuti ka geenide häirituse ja kromatiini immunosadestamise eksperimentide andmeid.

### 2.3.1 Oct4 ChIP-seq

*Oct4 ChIP-seq* andmestik sisaldab endas geenide Ensembl'i ID-sid, mille promootori piirkonnas on *OCT4* seondunud.

- Geeni Ensembl'i ID
- Geeni nimi

### 2.3.2 Oct4 knock-down experiment

*OCT4 expression down* andmestikus on nimekiri geenide Ensembl'i ID-dest, mis näitavad muutust ekspressioonis, kui *OCT4* regulatsiooni taset on vähendatud. Andmed on saadud geenide häirituse eksperimendiga 1.5.3.

- Geeni Ensembl'i ID
- Geeni nimi

[5] biit.cs.ut.ee/escd/info.html

## 2.4 Signaali vahendavate geenide leidmine

Töö eesmärgiks on automaatselt leida geenidevahelisi seoseid eelpool kirjeldatud suuremahulistest eksperimentaalsetest andmetest (GENCODE 2.2.1, ChIP-seq 2.2.4 ja DNase I 2.2.3).

Esimese sammuna signaalivahendajate leidmiseks tuleb kombineerida GENCODE andmestikus olevaid andmeid geenide võimalike regulaatorite ja aktiivsete promootorite kohta. Juhul kui geeni regulatoorses alas on näidatud transkriptsioonifaktori seondumist, siis peame seda võimalikuks signaalivahendajaks peamiselt regulaatorilt.

Teise sammuna tuleb geenikespressiooni andmetest eraldada huvipakkuva transkriptsioonifaktori poolt otseselt reguleeritud geenid. Alles jäänud geenidest tuleb eraldada need, mis on huvipakkuva faktori poolt kaudselt reguleeritud 2.2.5.

Kolmanda sammuna tuleb leida potentsiaalsed signaalivahendajad, mis on ise peamise regulaatori poolt otseselt mõjutatud. Kui selline kolmik leidub, kus huvipakkuv faktor mõjutab otseselt geeni X ja selle seondumine on leitud geeni Y promootori piirkonnas, siis saame geeni X pidada signaalivahendajaks huvipakkuvalt faktorilt.

### 2.4.1 Geenid ja nende ümbruses seondunud transkriptsioonifaktorite leidmine

Leidmaks geenide ümbrusesse seondunud transkriptsioonifaktorid, on esmalt vaja teada, kus genoomil paiknevad geenid. Selle saame teada GENCODE andmestikust, kus on antud, mis kromosoomil geen paikneb, geeni ID ning transkriptsiooni alguspunkti koordinaadid kromosoomil. Iga geeni jaoks tuleb määrata kaugus transkriptsiooni alguspunktist (TSS - ingl. k. *Transcription Start Site*), mille piirkonnas hakkame seonduvaid transkriptsioonifaktoreid otsima. Otsitava piirkonna suuruse valik on vaba (näiteks +/− 3000 aluspaari) ning antud regiooni jaoks on vaja ära märkida kõik seonduvad transkriptsioonifaktorid, mis leiduvad selles alas. Transkriptsioonifaktorite seondumise kohta on informatsioon ChIP-seq andmestikus, kust saab kätte seondumispiirkonna koordinaadid. Lisaks tuleb iga geeni ümbruse kohta kaardistada DNase I andmestikust saadaolevad regioonid, mis annab teavet selle kohta, kas DNA on üldse avatud ja kas sinna saavad TF-id seonduda. Kui on märgata transkriptsioonifaktori seondumist, siis tuleb arvutada ka TF-i seondumispiirkonna ülekatte % DNase I katsest saadud piirkonnaga.

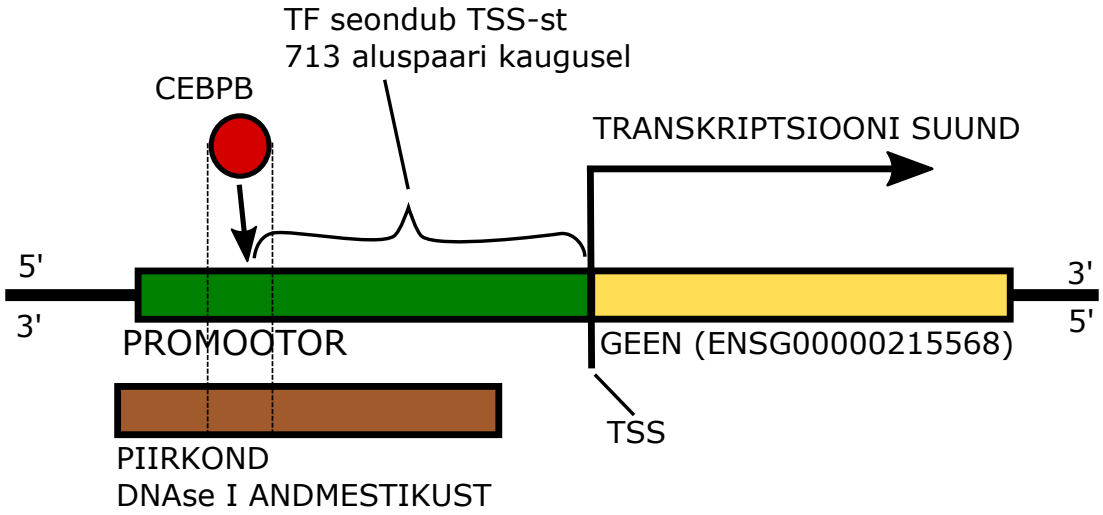
Esimese etapi jaoks on vaja luua andmestik järgmiste tulpadega, kus on iga transkriptsioonifaktori kohta järgmine informatsioon:

- Kromosoom
- Geeni nimi
- Geeni TSS koordinaat
- Rakuliin
- Transkriptsioonifaktori nimi
- Transkriptsioonifaktori keskpunkti kaugus geeni TSS koordinaadist
- DNase I ülekatte protsent juhul, kui see leidub

Esimese etapi tulemus on realiseeritud algoritmiga 1, parameetriteks on antud vahemik 3000 ja rakuliin H1-hESC. Selliste parameetritega loodud andmestikus on kokku 1641802 kirjet.) Järgnevalt on toodud näide andmestikust mingi geeni kohta:

| Veeru nimi | Veeru väärtus |
|---|---|
| Kromosoom | chr22 |
| Geeni Ensembl'i ID | ENSG00000215568 |
| TSS | 17449273 |
| Rakuliin | H1-hESC |
| TF-i nimi | CEBPB |
| Transkriptsioonifaktori keskpunkti kaugus geeni TSS koordinaadist | +713.0 |
| DNase I piirkonna ülekatte % | 100% |

**Tabel 1:** Näide esimese etapi tulemusest. Geeni ENSG00000215568 transkriptsiooni alguspunktist ülesvoolu (5' poole) paikneb 713 aluspaari kaugusel transkriptsioonifaktor nimega *CEBPB*. Samuti asub *CEBPB* täielikult piirkonnas, mis on saadud DNase I 2.2.3 andmestikust.




**Joonis 5:** Tabeli 1 andmed visuaalsel kujul. Rohelisega on tähistatud promootori ala kuhu transkriptsioonifaktor CEBPB seondub. Kaugus transkriptsiooni alguspunktist on väljendatud alsupaaride arvuga, mis jääb seondumispiirkonna keskpunkti ja TSS-i vahele. Pruuniga on tähistatud piirkond DNase I andmestikust, kuhu kõige tõenäolisemalt saab faktor seonduda, sest piirkond genoomil on avatud. Kollasega on tähistatud geen, mida hakatakse transkribeerima, sulgudes on toodud geeni Ensembl ID.

### 2.4.2 Geeni nimede tõlkimine ja kaudsete sihtmärkgeenide eraldamine

Teise etapi eesmärgiks on esimeses etapis loodud transkriptsioonifaktorite seondumissaitide (TFBS - ingl. k. *Transcription Factor Binding Sites*) failis Ensembl'i ID tõlkimine geeni nimeks 2.2.2. Nimede tõlkimine on vajalik, sest Ensembl ID on unikaalne identifikaator Ensembl'i andmebaasist ning geeni nime saab hiljem vajalikuks identifikaatoriks ümber tõlkida. Seejärel tulevad kasutusse lisaks üldistele andmetele kaks uut huvipakkuva transkriptsioonifaktori spetsiifilist andmestikku, mis on võetud ESCDb-st. Antud andmestike juures on kasutusel Python'i funktsioon *Sniffer()*[6], millega saab tuvastada, mille abil on andmed ridadel eraldatud, kuna sisendandmed ei pruugi alati samas formaadis

[6]https://docs.python.org/2/library/csv.html csv.Sniffer

olla. Esimeses andmestikus on huvipakkuva faktori poolt otseselt reguleeritud sihtmärgid 2.3.1, mis on saadud kromatiini immunosadestamise tulemusel, ning teiseks andmefailiks on geenide häirituse eksperimendi andmed 2.3.2, kus on huvipakkuva faktori poolt kaudselt mõjutatud geenid.

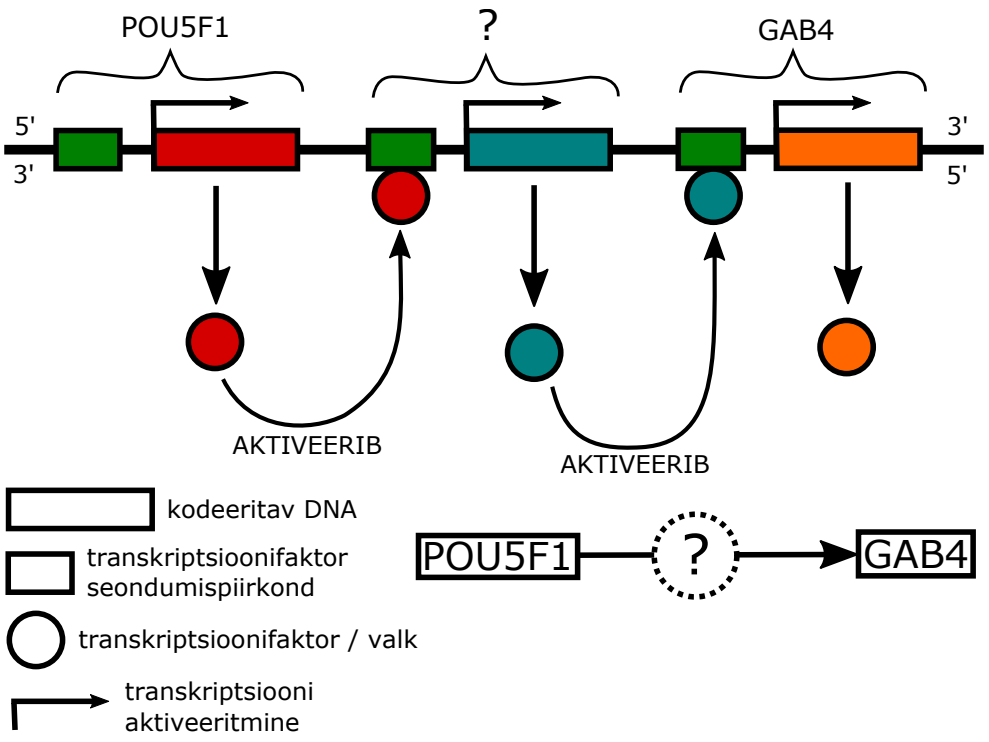
Teise etapi esimeses osas tõlgime 1. sammus loodud TFBS failis geeni Ensembl ID-d geeni nimedeks.

| Geeni Ensembl'i ID | Geeni nimi |
|---|---|
| ENSG00000215568 | GAB4 |

**Tabel 2:** Tõlkimine. Ensembl'i ID-le vastav geeni nimi.

Teise etapi teises sammus tuleb TFBS failist välja sorteerida otseselt mõjutatud geenid (andmestikus 2.3.1) ning samuti need geenid, mille ümbruses on huvipakkuv faktor seondunud.

Teise etapi viimases sammus tuleb allesjäänud geenide nimekirjas hoida alles ainult need, mis on huvipakkuva faktori poolt kaudselt mõjutatud (geenide häirituse eksperimendi andmestikus 2.3.2).




**Joonis 6:** Tundmatu signaalivahendaja. Geenide häirituse eksperimendist saame teada, et POU5F1 avaldab mõju GAB4'le, aga ta ei ole GAB4 otsene regulaator. Seda teame selle pärast, et töö teise etapi käigus eraldasime TFBS andmestikust kõik POU5F1 poolt otseselt reguleeritud sihtmärgid. Järelikult toimub regulatsioon läbi tundmatu signaalivahendaja.

Alles jääb andmestik, kus on huvipakkuva transkriptsioonifaktori kaudsed sihtmärkgeenid koos neid ümbritsevate transkriptsioonifaktoritega. Viimase etapi eesmärgiks ongi leida signaalivahendajad (joonisel 6 märgitud küsimärgiga).

### 2.4.3 Signaalivahendajate leidmine

Töö kolmanda sammuna leiamegi signaalivahendajad huvipakkuvalt transkriptsioonifaktorilt.

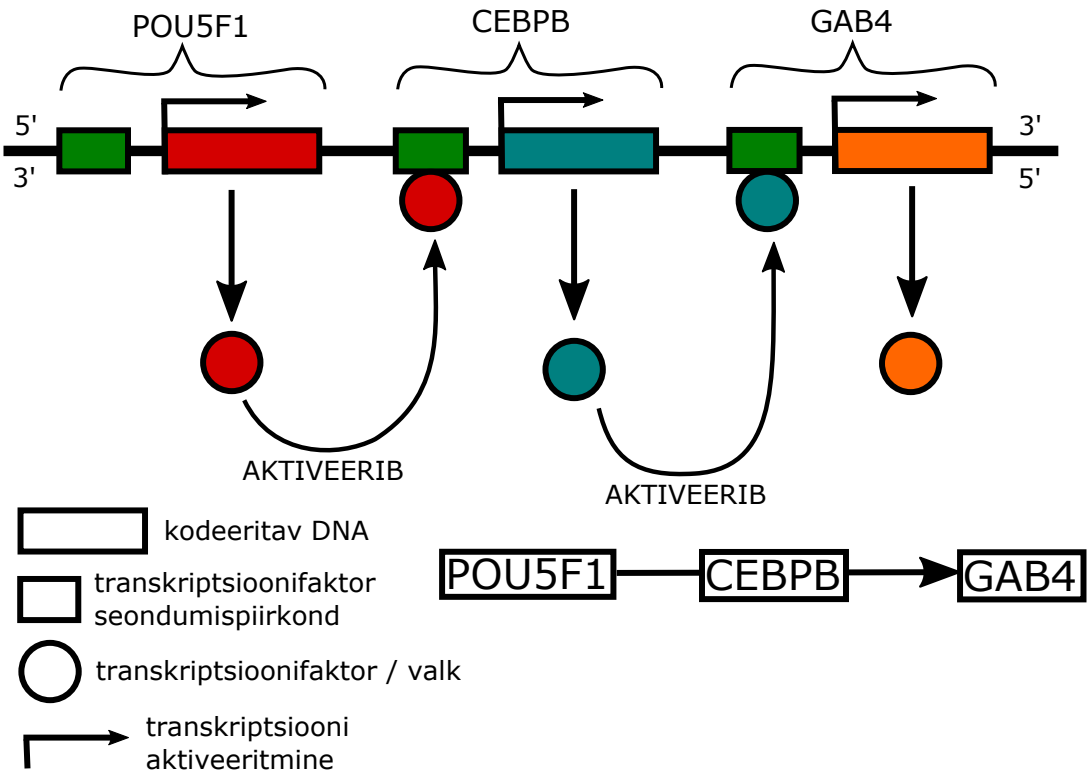
Kahe esimese etapi tulemusena on meil andmestik, kust on eraldatud huvipakkuva faktori teadaolevad otsesed sihtmärgid ESCDb andmebaasist ning samuti need geenid, mille ümbruses antud transkriptsioonifaktor seondunud on. Järele on jäänud ainult huvipakkuva faktori poolt kaudselt reguleeritud sihtmärgid ning nende geenide piirkonnas seondunud transkriptsioonifaktorid. Samuti on ESCDb andmebaasist olemas ka antud faktori poolt otseselt reguleeritud sihtmärkgeenide andmestik 2.3.1.

Signaalivahendajate leidmiseks tuleb neid kahte andmestikku kombineerida ning leida need geenid, mille ümbruses on märgatud huvipakkuva faktori poolt otseselt reguleeritud sihtmärgi seondumist.

Näide mingi geeni kohta pärast töö esimese kahe etapi läbimist:

| Veeru nimi | Veeru väärtus |
|---|---|
| Kromosoom | chr22 |
| Geeni nimi | GAB4 |
| TSS | 17449273 |
| Rakuliin | H1-hESC |
| TF-i nimi | CEBPB |
| Transkriptsioonifaktori keskpunkti kaugus geeni TSS koordinaadist | +713.0 |
| DNase I piirkonna ülekatte % | 100% |

**Tabel 3:** Geeni nimi on ID kaudu tõlgitud ja alles on jäänud ainult huvipakkuva faktori kaudsed sihtmärgid. See tähendab, et antud geen ei leidu otseste sihtmärkide seas (andmestikus 2.3.1). Antud tabeli näite puhul on vaja leida, kas TF nimega CEBPB leidub huvipakkuva transkriptsioonifaktori poolt otseselt reguleeritud sihtmärkide seas. Kui leidub, siis saame lugeda transkriptsioonifaktorit CEBPB üheks võimalikuks signaalivahendajaks peamiselt regulaatorilt geenile GAB4.




**Joonis 7:** Regulatoorse kolmiku näide. *POU5F1* avaldab mõju geenile *CEBPB*, mis omakorda mängib rolli *GAB4* geeni regulatsioonis. Antud juhul ongi üheks võimalikuks signaalivahendajaks *CEBPB*.

Näiteks kui leidub, et CEBPB on otseselt reguleeritud peamise regulaatori (antud juhul POU5F1) poolt ja loodud tabeli nr. 3 põhjal saame öelda, et CEBPB on seondunud geeni GAB4 ümbrusesse. Sellisel juhul loeme CEBPB-d üheks võimalikuks signaalivahendajaks geenilt POU5F1. Kirjeldatud olukord on visualiseeritud joonisel 7.

# 3 Signaalivahendajate automaatne leidmine

Signaalivahendajate automaatseks leidmiseks on selle töö raames loodud programm *Black Box Solver*.

## 3.1 Black Box Solver

Programm on kirjutatud keeles Python ning kasutatud on standardteeke:

- sys - Süsteemi spetsiifika
- codecs - Täpitähtede ja muude sümbolite encoding
- csv - Andmefailide sisselugemine
- os - Operatsioonisüsteemi spetsiifika
- time - Algoritmi jookmiseks kulunud aja mõõtmine
- configparser - Konfiguratsioonifailide töötlemiseks
- json - JSON kujul andmete töötlemiseks

### 3.1.1 Konfiguratsioon

Kasutajal on võimalik konfigureerida Black Box Solver'it oma soovide järgi, et saada tahetud tulemusi. Konfiguratsioon on failis *black_box_config.ini* ning vastavate parameetrite muutmiseks tuleb konfiguratsioonifail avada vabalt valitud tekstiredaktoris ning vastavaid parameetreid muuta.

```
[blackbox]
vahemik = 3000
rakuliin = H1-hESC
gencode = gencode.v19.TSS.notlow.gff
chipseq = chipseq_proximal.bed
dnasei = dnase_proximal.bed
geeninimed = ensg_genename.txt

[TFBSfile]
loodud_failid = []
```

**Kood 1:** Näide programmi konfiguratsioonifailist black_box_config.ini koos konfigureeritavate parameetritega.

Programmi käivitamisel kontrollitakse, kas konfiguratsioonifail eksisteerib. Kui see puudub, siis tekitatakse esialgne konfiguratsioon vaikeväärtustega (konfiguratsioon - 1).

Kasutaja poolt on määratavad järgmised parameetrid:

- blackbox
  - vahemik - Otsingupiirkond ehk kui kaugelt TSS-ist otsida aktiivseid promootoreid ja regulaatoreid (Täisarv)
  - rakuliin - rakuliini nimetus (Sõne tüüp)
  - gencode - GENCODE andmestik (Sõne tüüp)

  - chipseq - ChIP-seq andmestik (Sõne tüüp)
  - dnaseI - DNase I andmestik (Sõne tüüp)
  - geeninimed - ENSG geeni nimed tõlkimise jaoks (Sõne tüüp)

Programm ise täiendab vastavalt etteantud tööle järgmist parameetrit, kus hoitakse informatsiooni loodud TFBS failide kohta:

- TFBSfile
  - loodud_failid - loend loodud TFBS failidest kujul [[vahemik, rakuliin, faili nimi], ...]

### 3.1.2 Kasutamine

Programmi kasutamiseks on vajalik Python 3.x ning programmi tuleb käivitada käsurealt:

*python blackbox.py [huvipakkuva faktori nimi] [otsesed sihtmärgid] [häirituse eksperimendi andmed] [vahemik] [rakuliin]*

Programm võtab minimaalselt 3 argumenti:

1. Esimeseks argumendiks on huvipakkuva faktori nimi.
2. Teiseks argumendiks on huvipakkuva faktori või geeni poolt otseselt reguleeritud sihtmärgid ehk geenid, mille piirkonnas on märgatud antud faktori seondumist.
3. Kolmandaks argumendiks on huvipakkuva faktori kaudsed sihtmärgid ehk geenide häirituse eksperimendi tulemusena saadud andmed.
4. Neljas argument ei ole kohustuslik. Kui vahemik on määratud, siis programm kasutab seda, kui vahemikku ei ole määratud, võetakse vastav vaikeväärtus konfiguratsioonifailist 1.
5. Viies argument ei ole kohustuslik. Kui rakuliin on määratud, siis programm kasutab etteantud rakuliini. Kui ei ole määratud, siis rakuliini vaikeväärtus loetakse konfiguratsioonifailist 1.

## 3.2 Programmi töö käik

Programmi töö on jagatud loogiliselt nelja etappi:

1. Kontrollid
2. TFBS andmestiku lugemine/loomine
3. Signaalivahendajate leidmine ja andmestiku loomine
4. Suunatud graafi andmete loomine

### 3.2.1 Kontrollid

Esmalt kontrollitakse üle kõik programmi tööks vajalik.

1. Konfiguratsioonifail
    - Kui konfiguratsioonifaili ei leidu, loob programm uue konfiguratsiooni vaikeseadistustega (Konfiguratsioonifail 1).
    - Kui konfiguratsioonifail leidub, loeb programm konfiguratsiooni olemasolevast failist.
2. Programmi argumentide väärtused
    - Kontrollitakse, kas programm on saanud minimaalsed vajalikud argumendid.
    - Kontrollitakse, kas esimeseks argumendiks antud otseste sihtmärkide andmestik on olemas.
    - Kontrollitakse, kas teiseks argumendiks antud geeni häirituse eksperimendi andmestik on olemas.
3. Konfiguratsioonifailis kirjeldatud üldiste andmestike kontroll
    - Kontrollitakse, kas eksisteerib GENCODE andmestik 2.2.1.
    - Kontrollitakse, kas eksisteerib tõlkimiseks vajalik geeni nimede andmestik 2.2.2.
    - Kontrollitakse, kas eksisteerib DNase I andmestik 2.2.3.
    - Kontrollitakse, kas eksisteerib ChIP-seq andmestik 2.2.4.
4. Vahemiku ja rakuliini kontroll
    - Kontrollitakse kas on antud vahemik ja/või rakuliin.
    - Kui vahemik on antud, veendutakse, et tegemist on täisarvuga.
    - Kui vahemik puudub, loetakse väärtus konfiguratsioonifailist.
    - Kui rakuliin puudub, loetakse väärtus konfiguratsioonifailist.

### 3.2.2 TFBS andmestik

Kui programmi tööks vajalikud kontrollid on läbitud, jätkab programm TFBS andmestikuga. Esmalt programm vaatab, kas mõni vastavate parameetritega TFBS andmestik on juba loodud. Tehtud parameetritega andmestikke kontrollitakse konfiguratsioonifailist loodud_failid atribuudist.

1. Rakuliin
    - Kontrollitakse, kas on tehtud otsing antud rakuliini põhjal.
2. Vahemik
    - Kontrollitakse, kas on tehtud otsing antud vahemikuga. Vahemik võib olla väiksem või võrdne eelnevalt tehtud otsingust.

Kui antud parameetritega TFBS andmestik eksisteerib, loeb programm selle sisse. Andmestiku failinimi saadakse konfiguratsioonifaili loodud_failid parameetrist.

Kui antud parameetritega TFBS andmestikku ei leidu, siis hakkab programm etteantud sisend-parameetritega uut TFBS andmestikku looma. Uue andmestiku loomisel salvestatakse see kujul "TFBS-vahemik-rakuliin.txt". Samuti lisatakse konfiguratsioonifaili loodud_failid loendisse uus väärtus:
*[vahemik, rakuliin, "TFBS-vahemik-rakuliin.txt"]*

### 3.2.3 Signaalivahendajate andmestik

Eelviimase sammuna hakkab programm leidma ja koostama signaalivahendajate andmestikku. Kuna kõik kontrollid on teostatud (alapeatükk 3.2.1), siis saab programm jätkata tööd, nagu on kirjeldatud alapeatükis 2.4.

Programm informeerib kasutajat, mis etapp hetkel käsil on:

- *Alustan otsimist...*
- *Otsesed eemaldatud...*
- *Ainult kaudsed sihtmärgid alles, otsin signaalivahendajaid...*

Kui signaalivahendajad on leitud, salvestatakse tulemus faili *black_boxes_[huvipakkuva faktori nimi]_[vahemik]_[rakuliin].txt*. Samuti annab programm kasutajale uue andmestiku loomisest teada.

### 3.2.4 Suunatud graafi genereerimine

Viimase sammuna koostab programm suunatud graafi andmed, et teha visualiseeringuid ja edasisi võrgustike analüüse.

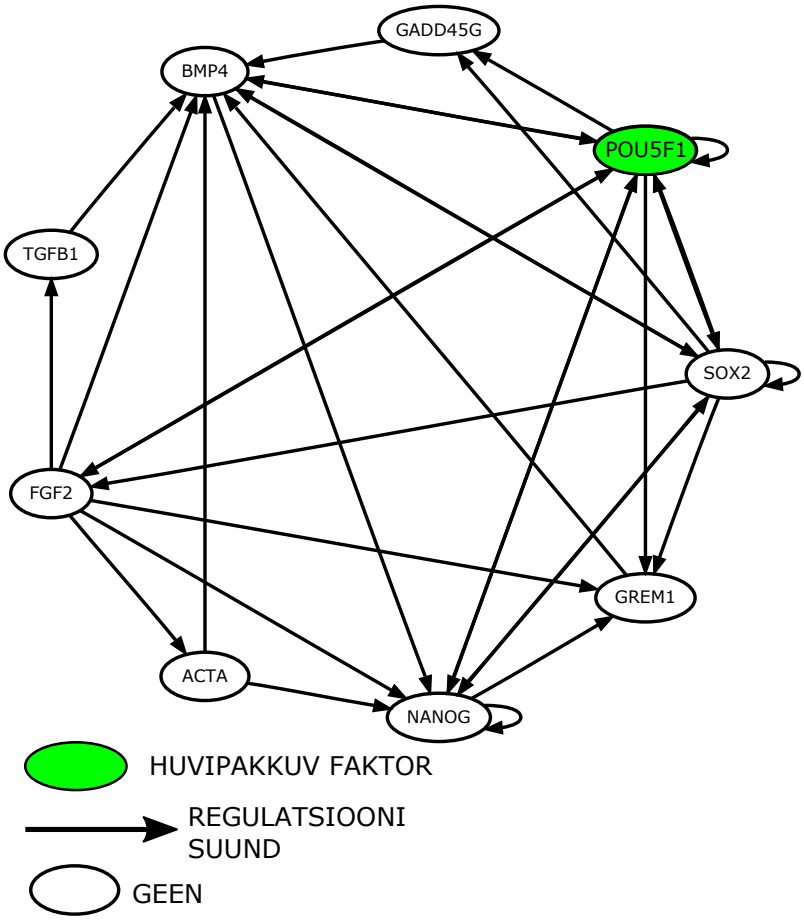
Graafi andmed genereeritakse kujul *[huvipakkuva faktori nimi] > [signaalivahendaja]; [signaalivahendaja] > [sihtmärk geen]*, kus > tähistab regulatsiooni suunda. Kuna regulatsiooni suund on teada, siis on meil tegemist suunatud graafiga. Sellisel kujul suunatud graafi on võimalik importida *GraphWeb*[7] [11] veebirakendusse, mis võimaldab kasutada erinevaid graafide jaoks mõeldud algoritme analüüsimaks signaalivahendajate võrgustikke. Visuaalsed võrgustikud on nähtavad järgmises, tulemuste peatükis.

[7] http://biit.cs.ut.ee/graphweb/

# 4 Tulemused

Toodud visualiseeringud on tehtud kirjanduses [5] avalikustatud teadaolevate seoste ja Black Box Solver'i genereeritud tulemuste kombineerimisel. Nii saame näidata, et Black Box Solver suudab teadaolevaid seoseid täiendada uute signaalivahendajatega. Graafide loomiseks on kasutatud *GraphWeb* veebirakendust. Visualiseerimisel on kasutatud on tugevalt sidusate tippude (ingl. k. *Strongly Connected Components*) algoritmi, et leida tugevalt sidusaid geene antud bioloogilises protsessis. Lisaks on valitud ainult POU5F1 võrgustiku ümbrus (ingl. k. *Network Neighbourhood*) kaugusega 2. See tähendab, et antud võrgustikus iga tipp ehk geen paikneb huvipakkuvalt faktorilt maksimaalselt 2 sammu kaugusel.

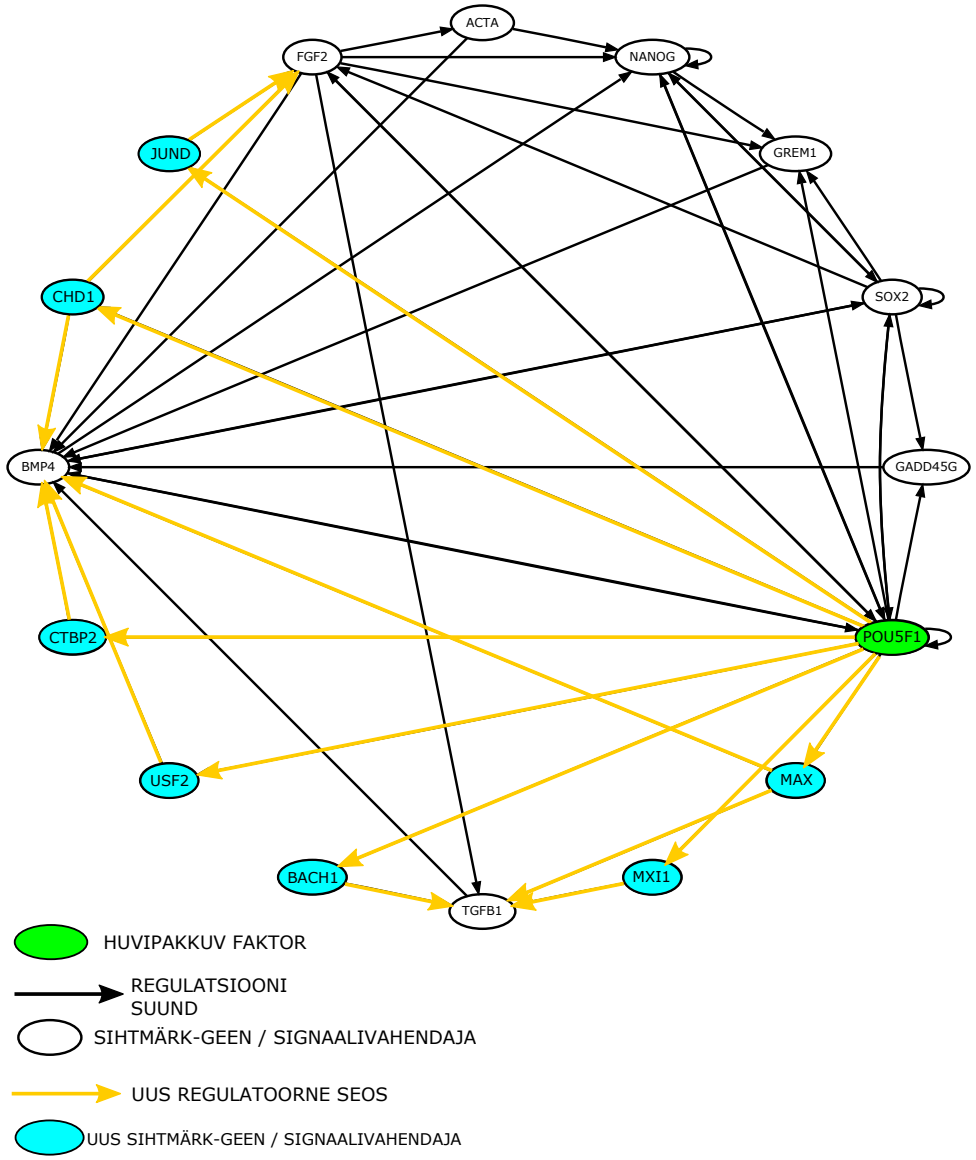
## 4.1 Tulemused kirjandusest




**Joonis 8:** Kirjandusest teadaolev geeni POU5F1 regulatoorne võrgustik. Rohelisega on märgistatud meile huvipakkuv faktor, mille signaalivahendajaid ja sihtmärk-geene otsime.

Antud töö käigus sai praktiline osa tehtud POU5F1 näitel. Joonisel 8 on toodud artikli [5] põhjal POU5F1 regulatoorne võrgustik, kus on näha antud faktori poolt mõjutatud sihtmärgid nii otseselt kui kaudselt.

## 4.2 Tulemused Black Box Solver'i ja kirjandusest saadud seoste kombineerimisel




**Joonis 9:** Kirjandusest teadaolevate seoste kombineerimine Black Box Solver'i tulemustega. Rohelisega on märgitud huvipakkuv faktor. Helesinisega on märgitud lisandunud regulaatorid. Kollaste nooltega on tähistatud lisandunud regulatoorsed seosed geenide vahel. Jooniselt on näha, kuidas Black Box Solver on välja pakkunud uusi võimalikke signaalivahendajaid, mis mängivad rolli POU5F1 regulatoorses võrgustikus.

Kasutades kirjandusest teadaolevaid seoseid ning kombineerides neid Black Box Solver'i tulemustega (parameetritega: faktori nimi POU5F1, OCT4 otsesed sihtmärgid 2.3.1, OCT4 häirituse eksperimendi andmed 2.3.2, vahemik suurusega 3000, rakuliin H1-hESC), saame täiendatud POU5F1 bioloogilise võrgustiku, mis lisab joonisel 8 kujutatud võrgustikule uusi signaalivahendajaid sihtmärk-geenidele. Bioloogiliste süsteemide keerukuse tõttu ei ole antud võrgustik täielik, kuid siiski on leitud uued hüpoteetilised seosed, mida saab edasi uurida.

# 5 Arutelu

Kromatiinisadestamine koos sekveneerimisega ja DNA mikrokiibiga on senini olnud transkriptsioonifaktorite seondumispiirkondade kaardistamisel väga edukas [6]. Sihtmärk-geenid kas aktiveeruvad antud transkriptsioonifaktori peale või ei reageeri üldse. Mikrokiibi tehnoloogiat kasutades on suudetud mõõta tegelikku geeni ekspressiooni huvipakkuva transkriptsioonifaktori häirituse korral, aga ei saa täpselt öelda, kas antud geen on otsene või kaudne sihtmärk.

Esmane samm mõistmaks geenide regulatsiooni, on leida tranksriptsioonifaktorite poolt otseselt mõjutatud geenid. Hiljutiste arengutega kromatiinisadestamise tehnoloogias saame kaardistada transkriptsioonifaktorite seondumispiirkondi tervel genoomil, aga täpset funktsionaalsust puhtalt selle järgi kirjeldada ei saa. Geenide häirituse eksperimendiga saame teada, kuidas reageerib geenide regulatoorne võrgustik, kui huvipakkuva geeni ekspressiooni taset märgatavalt tõsta või langetada. DNase I katsega saame teada piirkonnad, kus DNA on avatud ehk kuhu saab kõige tõenäolisemalt transkriptsioonifaktor seonduda.

Kombineerides neid tehnoloogiaid, saame palju täpsemini välja tuua otsesed ja kaudsed sihtmärgid ning nendega seondunud regulaatorid. Selle jaoks on loodud algoritmid, mis kasutavad kõiki tehnoloogiaid, et täpselt ära määrata otsesed ja kaudsed transkriptsioonifaktorite sihtmärgid [6] [8].

Olemasolevad veebitööriistad, nagu näiteks ChIP-Array [6], kasutavad kahekihilist andmete integreerimist (ChIP-* ja DNA mikrokiibi katsed). Antud töö raames loodud tööriist Black Box Solver kasutab samuti ChIP-*, DNA mikrokiibi tehnoloogiat, aga lisaks veel kombineerib ka DNase I andmestikust pärit avatud piirkondi, et lisada veelgi enam täpsust tulemustesse. Üldised andmestikud nagu GENCODE, ChIP-Seq ja DNase I saab ära defineerida konfiguratsioonifailis ning koheselt kasutama hakata. Huvipakkuva geeni kromatiini immunosadestamise ja häirituse eksperimendi andmed on pärit ESCDb andmebaasist, kus hoiatakse inimese ja hiire tüvirakkude andmeid ja nendega teostatud eksperimentide andmeid. Geeni nimede tõlkimiseks on kasutusel g:Convert tööriist, mis sünkroniseeritakse Ensembl'i baasiga iga paari kuu tagant, et hoida andmeid värsketena ja pakkuda kõige uuemaid tõlkeid. Samuti genereerib Black Box Solver bioloogilise võrgustiku jaoks suunatud graafi andmed, mida saab kasutada *GraphWeb* [11] veebirakendusega, et visuaalselt kuvada ja analüüsida geenide regulatoorseid võrgustikke 4.

# 6 Kokkuvõte

Selliseid kolmikuid ( [häiritud geen] → [signaalivahendaja] → [sihtmärkgeen] ) ja geeni funktsionaalset võrgustikku on vaja uurida, sest enamus haigused kujunevad inimesel mitme geeni koostoimel. Samuti on ka sellised haigused pärilikud, aga neil puudub kindel pärandusmuster. Kui võrrelda puhtalt lapsevanema ja lapse genoomi, siis ainult selle põhjal ei saa öelda, kas lapsele on haigus pärandunud ja kas on võimalik, et antud haigus lapses üldse avaldub. Küll aga kui me teame haiguse põhjustajaid ja kõiki vahelülisid, mis mängivad rolli geenides, mis on haigusega seotud, saame täpsemalt hinnata, kas haigus võib avalduda.

Esialgsed tulemused 4.2 on näidanud, et Black Box Solver on suutnud olemasolevatele tuntud seostele omalt poolt lisada hüpoteese, millised vahepealsed regulaatorid võiksid veel kuuluda huvipakkuva faktori regulaatorsesse võrgustikku. Selleks, et täielikult veenduda, kas esitatud hüpoteesid ka tõeks osutuvad, saab tõestada vaid katseliselt. Kuna katsete tegemine on nii aja kui ka ressursside kulukas, siis loodud tööriista eesmärgiks ongi püstitada hüpoteese ehk leida suurest geenide võrgustikust võimalikud signaalivahendajad ning edastada see info bioloogidele, kes saavad seda katseliselt tõestada. Sellega saame vähendada tehtavate katsete arvu ning hoida kokku aega.

Antud bakalaureusetöö juures oli huvitav kogu temaatika. Esiteks kokkupuude suuremahuliste andmetega, nende töötlmisega, kus algoritmide kestvus ei ole paar minutit, vaid hoopis paarkümmend minutit, ning saadud tulemuste hulk, mis küündis teatud tingimustel üle miljoni kirjeni. Teiseks töö bioloogiline pool ning tähtsus. Gümnaasiumi jooksul saadud teadmised andsid kõigest pealiskaudse ülevaate sellest, mis on geenid ja mis tähtsus on geeniregulatsioonil. Kui esmalt seisnes töö puhtalt andmete töötlemisel, puudus teadmine, miks selline andmete integreerimine on kasulik ning kuidas see kajastub reaalses maailmas. Tutvudes lähemalt bioloogilise taustaga, hakkasid andmed omama palju suuremat tähtsust ning see oli kogu töö juures kõige suurem motivatsioon. Teades, kuidas andmed on saadud ja mis on nende roll geeniregulatsioonis, tekkis kohe ka kõrgendatud huvi töö raames saadud tulemuste vastu. Samuti loodud algoritmide ja tarkvara suhtes, et programmi töös ei tekiks arusaamatusi.

# Viited


[1] MD Dr Ananya Mandal. *What is Gene Expression?* URL: http://www.news-medical.net/health/What-is-Gene-Expression.aspx.

[2] I. Dunham et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (2012), pp. 57–74.

[3] Nature Education. *Cell biology for seminars.* URL: http://www.nature.com/scitable/ebooks/cell-biology-for-seminars-14760004/contents.

[4] Nature Education. *The Information in DNA Is Decoded by Transcription.* URL: http://www.nature.com/scitable/topicpage/the-information-in-dna-is-decoded-by-6524808.

[5] Abhishek Garg Ying Wang Jaak Vilo Ioannis Xenarios Hedi Peterson Raed Abu Dawud and James Adjaye. *Qualitative modeling identifies IL-11 as a novel regulator in maintaining self-renewal in human pluripotent stem cells.* URL: http://journal.frontiersin.org/article/10.3389/fphys.2013.00303/full.

[6] Panwen Wang Michael Q. Zhang Jing Qin Mulin Jun Li and Junwen Wang. *ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor.* URL: http://nar.oxfordjournals.org/content/39/suppl_2/W430.full.

[7] Marc Jung et al. "A data integration approach to mapping OCT4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells". In: *PLOS One* 5.5 (2010), e10709.

[8] Kevin P. White Roger Sciammas Mark Maienschein-Cline Jie Zhou and Aaron R. Dinner. *Discovering transcription factor regulatory targets using gene expression and binding data.* URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3259433/.

[9] Craig L. Peterson Michael F. Carey and Stephen T. Smale. *Chromatin Immunoprecipitation (ChIP).* URL: http://cshprotocols.cshlp.org/content/2009/9/pdb.prot5279.full.

[10] J. Reimand, T. Arak, and J. Vilo. "g:Profiler–a web server for functional interpretation of gene lists (2011 update)". In: *Nucleic Acids Research* 39.Web Server issue (2011), W307–315.

[11] J. Reimand et al. "GraphWeb: mining heterogeneous biological networks for gene modules with functional significance". In: *Nucleic Acids Research* 36.Web Server issue (2008), W452–459.

[12] Nick True Wei Wang Ron X. Yu Jie Liu. *Identification of Direct Target Genes Using Joint Sequence and Expression Likelihood with Application to DAF-16.* URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0001821.

[13] How HeLa Cells Work. *Shanna Freeman.* URL: http://science.howstuffworks.com/life/cellular-microscopic/hela-cell.htm.

# Lisa I - Algoritmid

## Geenide ja nende ümbruses seondunud transkriptsioonifaktorite leidmise algoritm

**Algorithm 1:** Step 1 - TFBS algoritm

**Input**: ENCODE 2.2.1, chipseq 2.2.4, DNase I 2.2.3
**Result**: Geeni transkriptsioonifaktorite seondumispiirkondade tabel koos DNase I piirkonna ülekattuvusega

```
1  kaugusTSSist = 3000
2  foreach kromosoom IN gencode.v19 do
3      Võta nimekiri geenidest, mis asetsevad antud kromosoomil.
4      Võta nimekiri transkriptsioonifaktoritest koos nende seondumispiirkondadega
       antud kromosoomil andmestikust chipseq-proximal
5      foreach geen IN praeguses kromosoomis do
6          Võta praeguse geeni TSS koordinaat.
7          Leia DNase I regioonid, mis asetsevad +/- aluspaari kaugusel TSS-st
           andmestikust dnase-proximal
8          foreach DNase I regioon IN dnase-proximal do
9              if DNase I regiooni algus >= TSS koordinaat - kaugusTSSist VÕI
               DNase I regiooni lõpp <= TSS koordinaat + kaugusTSSist then
10                 Jätame piirkonna meelde ja salvestame nimekirja;
11         foreach transkriptsioonifaktor IN antud kromosoom do
12             Arvuta transkriptsioonifaktori seondumispiirkonna keskpunkti
               koordinaat (Algoritm 2)
13             if TSS koordinaat - kaugusTSSist <= Transkriptsioonifaktori
               seondumispiirkonna keskpunkti koordinaat <= TSS koordinaat +
               kaugusTSSist then
14                 foreach DNase I piirkond geeni +/- aluspaari raadiuses do
15                     Leia transkriptsioonifaktori seondumispiirkonna ja DNase I
                       regiooni kattumise % (Algoritm 3)
16                 Kromosoom | Geen | TSS | Rakuliin | TFname | Kaugus | DNase I
                   Ülekatte %
17             else
18                 Võta ette järgmine transkriptsioonifaktor.
19 return Nimekiri geenidest ja nende piirkonnas olevatest transkriptsioonifaktoritest
   koos DNase I piirkonna ülekatte %-ga
```

## Seondumispiirkonna keskpunkti koordinaadi arvutamine

**Algorithm 2:** Transkriptsioonifaktori seondumispiirkonna keskpunkti koordinaadi arvutamine

**Input**: Seondumispiirkonna alguspunkti koordinaat $\mathbf{X}$, Seondumispiirkonna lõpp-punkti koordinaat $\mathbf{Y}$

**Result**: Koordinaatide keskpunkt

1

$$\mathbf{X} + (\mathbf{Y} - \mathbf{X})/2$$

## DNase I andmestiku piirkonna ülekatte arvutamine

**Algorithm 3:** Transkriptsioonifaktori seondumispiirkonna ja DNase I katsest saadud piirkonna ülekattumise %

**Input**: DNase I katse piirkonna algus ($X_1$); DNase I katse piirkonna lõpp ($X_2$), TF-i seondumispiirkonna algus ($Y_1$), TF-i seondumispiirkonna lõpp ($Y_2$)

**Result**: Mitu % katab TF-i seondumispiirkond DNase I katsest saadud piirkonda

```
1  if X1 ⇐ Y1 ⇐ X2 AND X1 ⇐ Y2 ⇐ X2 then
2  |  return 100%
3  else
4  |  if Y1 < X1 AND X2 < Y2 then
5  |  |  return 100 − (((Y2 − Y1 ∗ 100)/X2 − X1) − 100)
6  |  else
7  |  |  if Y1 < X1 AND X1 ⇐ Y2 ⇐ X2 then
8  |  |  |  return (Y2 − X1 ∗ 100) − Y2 − Y1
9  |  |  else
10 |  |  |  if X1 ⇐ Y1 ⇐ X2 AND X2 < Y2 then
11 |  |  |  |  return (X2 − Y1 ∗ 100)/X2 − X1
12 |  |  |  else
13 |  |  |  |  return 0
```

# Lisa II - Lähtekood

Lähtekood on kaasas *black_ box_ solver.zip* failis. Fail sisaldab endas:

- blackbox.py
  - Konfiguratsioonifaili lugemine/loomine
  - Üldiste andmete kontroll
  - Otseselt mõjutatud geenide andmete kontroll
  - Geeni häirituse eksperimendi andmete kontroll
  - Vahemiku ja rakuliini kontroll
- blackboxutil.py
  - TFBS faili kontroll
  - Signaalivahendajate leidmise algoritm
- TFBSUtil.py
  - GENCODE andmestiku lugemine
  - ChIP-Seq andmestiku lugemine
  - DNase I katse andmestiku lugemine
  - Geeni nimede tõlgete lugemine
  - TFBS andmestiku loomise algoritm (Algoritm 1)
  - TF-i seondumispiirkonna keskpunkti arvutamine (Algoritm 2)
  - Ülekatte kalkulaator (Algoritm 3)
- property_file.py
  - Vaikeväärtustega konfiguratsioonifaili loomine
- kolmikud.py
  - Signaalivahendajate andmestiku lugemine
  - Unikaalsete kolmikute loomine (korduste vältimine)
  - *GraphWeb'ile* sobilike sisendandmete genereerimine

# Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Andreas Ellervee (sünnikuupäev: 29 aprill 1993),

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose:
Geeniregulatsiooni signaali vahendavate geenide automaatne leidmine suuremahulistest eksperimentaalsetest andmetest,

mille juhendajad on: Hedi Peterson ja Elena Sügis,

1.1 reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2 üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartu, 14.05.2015