T A R T U Ü L I K O O L
BIOLOOGIA-GEOGRAAFIA TEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
Bioinformaatika õppetool

Hedi Peterson

Geeniregulatsiooni andmebaas BiGeR

Keskastme lõputöö

Juhendaja: Jaak Vilo, PhD

Tartu 2003

Sisukord

Lühendid

A	Adenine	Adeniin
API	Application Programming Interface	Rakendusliides
Arg	Arginine	Arginiin
C	Cytosine	Tsütosiin
Cys	Cysteine	Tsüsteiin
DBI	Database Interface	Andmebaasi kasutajaliides
DDL	Data Definition Language	Andmete defineerimiskeel
DML	Data Manipulation Language	Andmete manipuleerimiskeel
DNA	Deoxyribonucleic acid	Desoksüribonukleiinhape
DQL	Data Query Language	Andmete pärimiskeel
G	Guanine	Guaniin
Gly	Glycine	Glütsiin
His	Histidine	Histidiin
HMM	Hidden Markov Model	Peidetud Markovi Mudel
IUPAC	International Union of Pure and Applied Chemistry	Rahvusvaheline Puhta- ja Rakenduskeemia Liit
Lys	Lysine	Lüsiin
mRNA	messenger–RNA	matriits–RNA
ORF	Open Reading Frame	Avatud lugemisraam
PCR	Polymerase Chain Reaction	Polümeraasi ahelreaktsioon
Phe	Phenylalanine	Fenüülalaniin
Pro	Proline	Proliin
PROSITE	Database of protein families and domains	Valgu perekondade ja domäänide andmebaas
PSSM	Position specific score matrix	Positsioonispetsiifiline skoorimaatriks
PWM	Position weight matrix	Positsiooni kaalumaatriks
RDBMS	Relational Database Management System	Relatsiooniline andmebaaside juhtimise süsteem
RNA	Ribonucleic acid	Ribonukleiinhape
SQL	Structured Query Language	Struktuurpäringukeel
T	Thymine	Tümiin
TSS	Transcription start site	Transkriptsiooni algussait
Tyr	Tyrosine	Türosiin
U	Uracil	Uratsiil

Sissejuhatus

Viimase kümnendi jooksul toimunud hüppeline areng erinevate organismide genoomide sekveneerimises on pannud aluse bioinformaatika tormilisele edasiminekule. Üha enam otsitakse DNAst bioloogiliselt olulisi signaale, toodetakse *in silico* ja *in vitro* uusi eksperimentaalseid ja ennustuslikke andmeid ning luuakse andmebaase saadud info esitamiseks. Selliste andmebaaside loomine on andnud võimaluse geeniregulatsiooni mehhanismide modelleerimiseks ja mõistmiseks, mis on tänapäeva molekulaarbioloogia suurimaid väljakutseid. Siiani on olulisel kohal pärmi *Saccharomyces cerevisiae* kui mudelorganismi uurimine ning kasutamine suuremahuliste analüüsiprotsesside väljatöötamiseks. Kuigi on loodud mitmeid erinevaid andmebaase geeniregulatsiooni andmete haldamiseks ja esitamiseks puudus senini võimalus kompleksek andmete päringuks.

Käesolev töö koosneb kahest peamisest osast – teoreetilisest, kirjanduse põhisest ülevaatest ning praktilisest, toimiva andmebaasi esmase prototüübi loomisest ning selle andmebaasi kirjeldusest käesolevas töös.

Teoreetilises osas esitatakse bioloogiliste signaaljärjestuste erinevaid leidmis-, esitus- ja analüüsimeetodeid ning iseloomustatakse bioloogiliste andmebaaside modelleerimist.

Eksperimentaalses osas käsitletakse *S. cerevisiae* transkriptsioonifaktorite seondumissaite kirjeldava ning geeniregulatsiooni modelleerimist võimaldava andmebaasi valmimise erinevaid etappe. Andmebaasi struktuurist antakse ülevaade skemaatiliste jooniste ning olemite ja atribuutide kirjeldamise kaudu. Andmebaasi funktsionaalsusest annab ülevaate kasutusjuhtude ning näidispäringute kirjeldamine.

Osa I

Teoreetiline osa

Peatükk 1

Kirjanduse ülevaade

Käesolevas peatükis käsitletakse pikemalt geeniregulatsiooni kontrolli peamist allikat, transkriptsiooni. Tuuakse lühike ülevaade geeniregulatsioonist transkriptsiooni tasemel, samuti käsitletakse transkriptsiooni molekulaarset kontrolli. Olulisel kohal on ka transkriptsioonifaktorite iseloomustamine ning transkriptsioonifaktorite seostumissaitide lühiiseloomustus. Peatüki viimases osas tuuakse ülevaade bioloogilistest andmebaasidest.

1.1 Geeniregulatsioon eukarüootsetes organismides

Eukarüootsete organismide kompleksne geeniregulatsioon on mehhanism, mis koosneb mitmetest eri etappidest: transkriptsioon, transkriptsioonijärgne RNA protsessimine, mRNA stabiilsuse ehk eluea kontroll ning translatsioon. Geeniregulatsiooni teevad keerukaks rakkude spetsiifilisus, pidev reageering keskkonna mõjudele ning eukarüootsete organismide üldine komplekssus. Geeniregulatsiooni peamiseks ülesandeks on organismis vajaminevate valkude õigeaegne ning õiges koguses ekspresseerimine. See on saavutatav otseste ja kaudsete transkriptsioonifaktorite erineva aktiivsusega või erineva afiinsusega RNA polümeraasile. Antud muutused tulenevad regulaatorvalkude seostumissaitide erinevusest.

Eukarüootsete organismide hulkrakulisus ning geenide avaldumine erinevates rakutsükli etappides, organismi erinevatel arenguetappidel ning erinevates kudedes teeb keerukaks transkriptsiooniregulatsiooni kontrolli ja nõuab tugevat kontrollsüsteemi, mis võimaldaks määrata vajalike geenide ekspresseerumise.

Geenide avaldumine on määratud ruumiliselt, ajaliselt ning keskkonna muutuste poolt. Geeniregulatsioon on kontrollitud mitmel erineval viisil ja

tasemel. Geenide avaldumine on kontrollitud esmalt transkriptsiooni tasemel ning seejärel valgu sünteesi kaudu. Eukarüootidel toimub enne translatsiooni RNA protsessimine, mille käigus RNA 5' otsa lisatakse cap–struktuur ning 3'otsa polü–A saba. Toimub ka intronite kõrvaldamine ehk splaissing. Nii transkriptsioon kui RNA modifitseerimine ja splaissimine toimuvad rakutuumas.

Eukarüootne geeniregulatsioon transkriptsiooni tasemel

Transkriptsioon on RNA süntees DNA kodeerivalt ahelalt. RNA sünteesi kontroll on geeni aktiivsuse regulatsiooni põhiline tase. Transkriptsiooni viib läbi DNA–sõltuv RNA polümeraas.

Transkriptsiooni kontrolli eukarüootsetes organismides teeb keerukaks keskkonnasignaalide komplitseeritud liikumine raku pinnalt tuuma. Seepärast peab eukarüootsetes organismides toimima sisemine signaalsüsteem, mis võimaldab transkriptsiooni regulatsiooni. Lisaks seab eukarüootsete organismide hulkrakulisus piirangud geeniregulatsioonile, mis peab suutma signaale saata läbi rakukihtide sihtmärgini, kus geene ekspresseeritakse. Geeniregulatsiooni kontrolliks on oluline nii rakusisene kui ka rakkudevaheline signaalsüsteem. Transkriptsiooni regulatsioon toimub valk–valk ja valk–DNA interaktsioonide kaudu. DNA spetsiifilistele regioonidele seostuvad positiivsed ja negatiivsed regulaatorvalgud ehk transkriptsioonifaktorid, mis stimuleerivad või inhibeerivad transkriptsiooni.

RNA splaissing

Valdav osa kõrgemate eukarüootsete organismide avatud lugemisraame (ORF-e) sisaldab introneid. Transleeritava mRNA saamiseks toimub splaissosoomides splaissimine, mille käigus lõigatakse intronid välja ja eksonid ühendatakse. Mitmete intronite esinemine geenis võimaldab alternatiivset splaissingut, mille käigus toimub intronite kõrvaldamine kas eraldi või koos eksonitega. Nii moodustuvad ühest geenist mitmed mRNA molekulid, mis kodeerivad erinevaid valke. Oluline on märkida, et erinevalt kõrgematest eukarüootidest on *S. cerevisiae* ORF-idest vaid 4%–l leitud introneid (Davis *et al.* 2000).

mRNA stabiilsuse tsütoplasmaatiline kontroll

Valgusüntees toimub tsütoplasmas. Translatsioon mRNA molekulilt toimub kuni molekuli degradeerumiseni. mRNA degradatsiooni kiirus määrab selle, kui kaua valku sünteesitakse. Lühikese degradatsiooniajaga molekulidelt toimub vähene valgu süntees ning seega tuleb geeni pidevalt ekspresseerida,

et oleks võimalik valku sünteesida. Valkude ajutine süntees on regulatoorse tähtsusega, võimaldades valgu sünteesi vaid teataval arenguetapil. Antud mehhanism on oluline, sest sageli on organismil vaid hetkeliselt vaja valku, mille pidev ekspresseerumine on kahjuliku mõjuga. mRNA eluiga mõjutavad 3' polü–A saba olemasolu ja pikkus ning 3' mittetransleeritavad regioonid (Cao & Parker 2001).

Eukarüootsete geenide transkriptsiooni molekulaarne kontroll

Transkriptsiooni kontrollis osalevad DNA järjestused, mille ülesanneteks on võimendada ning vaigistada transkriptsiooni. Transkriptsiooni initsiatsioon toimub geeni promooterregioonilt, mille tunneb ära RNA polümeraas. RNA polümeraasi seondumiseks promooteralale on vajalik eelnev basaalsete transkriptsioonifaktorite seondumine. Eukarüootsete geenide transkriptsiooni kontrollivad ka spetsiaalsed transkriptsioonifaktorid— võimendajad ja vaigistajad. Transkriptsioonifaktorid seostuvad regulatoorsetele DNA järjestustele ning stimuleerivad või pärsivad transkriptsiooni. Võimendajaid eristavad promooteritest järgnevad tunnused:

- võivad toimida pika vahemaa tagant
- mõju ei sõltu orientatsioonist
- mõju ei sõltu asukohast ehk võivad paikneda geenist eespool, tagapool või intronites.

Võimendajale ja promooterregioonile seondunud transkriptsioonifaktorid satuvad füüsilisse kontakti DNA lingude moodustumisel. Nende kokkusattumisel muutub transkriptsioonikompleks aktiivseks.

1.1.1 Transkriptsioonifaktorid

Geenide transkribeerimine on peamiselt seotud DNA–ga interakteeruvate valkude—transkriptsioonifaktorite, äratundmise ning DNA–le seostumisega. Transkriptsioonifaktorid seostuvad spetsiifilistele lühikestele DNA järjestusmotiividele geenide *cis*-regulatoorses piirkonnas. Keskkonnasignaalidest ning arenguetappidest sõltuvalt mõjutavad transkriptsioonifaktorid geenide aktivatsiooni ning repressiooni. Transkriptsioonifaktorite seostumist spetsiifilisele DNA alamjärjestusele on võimalik täpselt määrata bioloogiliste analüüsidega ning seda käsitletakse peatükis 2.1.

Transkriptsioonifaktorid on aktiivsena dimeerses ja inaktiivsena monomeerses vormis. Sõltuvalt dimerisatsioonist toimub DNA–ga seostumine.

DNA–ga seostuvaid valke iseloomustavad kolm põhilist tunnust (www.whatislife.com):

- Suur vagu, mis on valkude seostumissait α-heeliksites, on 12 Å lai ja 8 Å sügav.
- Väike vagu on 5 Å lai ja 8 Å sügav ning on liiga kitsas, et sobituda kogu α-heeliksiga. Väike vagu tuntakse ära TATA boksiga seonduvate valkude beeta struktuuri poolt.
- Järjestusspetsiifilised DNA-ga seostuvad valgud ei paiska üldjuhul segi DNA aluspaare, kuid moonutavad selgroo struktuuri väänates kaksikheeliksit.

Transkriptsioonifaktorite struktuursed motiivid

Transkriptsioonifaktorid sisaldavad mitmeid erinevaid struktuurseid motiive, mis seostuvad spetsiifilisele DNA järjestusele. Eukarüootsetel transkriptsioonifaktoritel DNA seostumisdomäänides olevad α– heeliksid on asetunud nii, et nad liiguksid DNA suurde vakku valgu ja DNA ühinemisel spetsiifiliste vesiniksidemetega. Sageli jaotatakse transkriptsioonifaktoreid nende DNA seostumisdomääni tüübi järgi. Enamusel DNA seostumisdomäänidel on iseloomulik aminohapete konsensusjärjestus. Seega on võimalik uued kirjeldatud transkriptsioonifaktorid jagada pärast geenide või cDNA sekveneerimist vastavasse domäänitüüpi. Peamised eukarüootsete transkriptsioonifaktorite struktuursed domäänid on tsink-sõrm, leutsiinilukk ja heeliks-silmus-heeliks. Käesolevas töös kirjeldatakse kõige detailsemalt enim esinevat tsink-sõrme motiivi ning antakse lühike ülevaade ka leutsiiniluku ning heeliks-silmus-heeliksi domäänidest (www.whatislife.com) põhjal.

Tsink-sõrm

Tsink–sõrm esineb kolmel erineval struktuursel kujul:

- Cys_2–His_2
- Cys_4
- Cys_6

Cys_2–His_2 tsink–sõrme domään on enim esinev DNA seostumisdomään eukarüootsetes transkriptsioonifaktorites. Koosneb kahest muutumatust Cys ja His jäägist. Zn^{++} on tetraeedriliselt seotud. Domääni iseloomustab (Tyr,Phe)-X-Cys-X_{2-4}-Cys-X_3-Phe-X_5-Leu-X_2-His-X_{3-5}-His motiiv. Vähem kui 50 aminohappelised järjestused ei voltu automaatselt.

Tsink–sõrme sekundaarstruktuur koosneb kahest antiparalleelsest kaheahelalisest β–lehest ning ühest α–heeliksist. Hüdrofoobne tuum koosneb kolmest hüdrofoobsest aminohappest. Tsink–sõrmed moodustavad omavahel 2–st kuni 37–st sõrmest koosnevaid tandeemseid kogumeid.

Tsingi ülesanne on väikse lingu struktuuri stabiliseerimine hüdrofoobsete tuumajääkide asemel. Iga tsinksõrm interakteerub konformatsiooniliselt ühtmoodi, moodustades edukaid kolme aluspaarilisi segmente kaksikheeliksi suures vaos. Valgu ja DNA interaktsioon on määratud kahe faktori poolt.

Esiteks vesiniksidemete interaktsioonid α–heeliksi ja DNA segmendi vahel, enamjaolt arginiini jäägi ja guaniini aluste vahel. Teiseks vesiniksidemete interaktsioonid DNA fosfaadi selgrooga, peamiselt arginiini ja histidiiniga.

Cys_4 domääni puhul tekivad 80 aminohappest koosnevad domäänid, mis koosnevad kahest "sõrmest". Esimene üksus seostub DNA–ga ja teine võimaldab kahe identse retseptormolekuli dimerisatsiooni. Sõrm koosneb ebaregulaarsest lingust ja heeliksist.

Cys_6 motiiv on binukleaarne tsinksõrm, kaks tsingi iooni on seotud tetraeedriliselt asetunud kuue Cys jäägiga. Iga Cys jääk on seotud kahe metalliiooniga. Domään koosneb 6st muutumatust Cys jäägist, mis on seotud suvalise aminohappega(X). Sellist motiivi iseloomustab Cys–X_2–Cys–X_6–Cys–X_6 Cys–X_2–Cys–X_6–Cys. Cys vahel asuvate aminohapete esinemise pikkused on rangelt konserveerunud. Domäänil on tetraeedriline asetus ja seostub DNA–ga sümmeetrilise dimeerina.

Selline struktuurne mitmekesisus on vajalik spetsiifiliste geenijärjestuste äratundmiseks ehk transkriptsioonifaktorite geenispetsiifilisuseks. Oluline on märkida, et mitte ühelgi juhul ei osale tsink ise seostumisinteraktsioonides vaid on koordinatiivseks aatomiks. (www.whatislife.com).

Leutsiinilukk

Umbes 30 aluspaariline motiiv koosneb kahest amfipaatsest heeliksist, mis interakteeruvad omavahel andes vasakpöördelise superspiraali sekundaarstruktuuri. Pro ja Gly jäägid motiivis puuduvad oma heeliksit lõhkuva iseloomu tõttu, palju on Arg ja Lys. Leutsiinilukk koosneb kahest antiparaleelselt seotud α–heeliksist, mida ühendavad regulaarselt asetsevad leutsiinid. α–heeliksis on iga seitsmes jääk leutsiin. Enamus leutsiinilukkudes asuvaid heelikseid esindavad seitsmeosalist järjestust, kus esimene ja neljas liige on hüdrofoobsed aminohapped ja ülejäänud hüdrofiilsed. Leutsiiniluku motiivid võivad vahendada kas homo- või heterodimeerset vormi. Leutsiiniluku motiiv ei ole DNA-ga seonduv osa heeliksist (www.whatislife.com).

Heeliks-silmus-heeliks

Heeliks-silmus-heeliks motiiv koosneb kahest α-helikaalsest regioonist, mida ühendab varieeruva pikkusega regioon, mis moodustab kahe heeliksi vahele silmuse. Heeliks-silmus-heeliks tüüpi motiiv on vajalik valkude seostumiseks

kahekordse telje sümmeetriaga järjestuse elementidega. Antud motiiv on vajalik valgu homo- ja heterodimerisatsiooniks (www.whatislife.com).

1.1.2 Transkriptsioonifaktorite seondumissaidid

Tervete genoomide kättesaadavus on motiveerinud teadlasi uurima transkriptsioonifaktorite seondumissaite *in silico* analüüsi abil. Seostumissaitide uurimine hõlmab kaht suurt ülesannet. *In vitro* on võimalik DNA seostumisspetsiifikat kindlaks teha DNaas I jalajäljega(*DNase I footprinting*) ja elektromobiilse nihke analüüsiga(*shift assay*) (Qiu 2003).

In silico transkriptsioonifaktorite seondumissaite ennustades on kaks peamist ülesande püstitust. Esimene ülesanne on ennustada tõenäosuslikke seostumissaite juba tuntud transkriptsioonifaktorile üle genoomi. Sellisel puhul kasutatakse teadaolevaid näiteid ning nende põhjal genereeritud mudeliga otsitakse genoomist uusi mudelile vastavaid võimalikke esinemisi.

Teine ülesanne on teadmata transkriptsioonifaktori seondumissaiti leida tõenäoliselt sama faktori poolt reguleeritud geenide ülesvoolu (*upstream*) järjestustest transkriptsioonifaktori seondumismotiiv ning seejärel leida saadud motiivi esinemised ka mujal genoomis.

Mõlemad ülesandepüstitused eeldavad transkriptsioonifaktorite seondumissaitides esineva motiivi kirjeldamist ja iseloomustamist. Transkriptsioonifaktorite seondumissaidid on 5-30 aluspaari pikad, seejuures enamik jääb vahemikku 5-16 aluspaari (Zhu & Zhang 1999; Vilo 2002; Qiu 2003). Seondumissaidi pikkus on sõltuv eksperimentaalsest meetodist, millega sait leiti (Zhu & Zhang 1999). Üherakulises organismis asuvad enamus regulatoorseid elemente 200–500 aluspaari ORFi 5' otsast ülesvoolu (Qiu 2003). Kuigi seostumissaidid on küllaltki konserveerunud, osutavad nad mõningasele varieeruvusele. Järjestuse motiiv peab esindama mitmeid lubatud seostumissaidi alamjärjestusi.

Regulaatorvalkute puhul on muutlikkus transkriptsiooni oluliseks kontrollsüsteemiks (Stormo 2000).

Enamkasutavateks seondumissaitide esitusviisideks on oligonukleotiidid, regulaaravaldised, maatriksid ning konsensusjärjestused. Neist motiividest tuleb pikemalt juttu peatükis 2.3.

1.2 Bioloogilised andmebaasid

1.2.1 Bioloogiliste andmebaaside vajadused

Maksimaalse kasu saamiseks bioloogiliste andmete rohkusest, tuleb need andmed siduda omavahel ühtseks tervikuks ning esitada kujul, mis võimaldaks teostada nii lihtsaid kui keerukamaid komplekspäringuid. Samuti on oluline esitada andmeid terviklahendusena (Birney, Clamp, & Hubbard 2002). Tervete genoomide kättesaadavus loob laiemad võimalused uurimaks bioloogiat kui tervikut. Järjestuste ulatuslik sekveneerimine on aidanud määratleda probleemide piirjooni ning pannud aluse edasistele uuringutele, mille täitmiseks on vaja uute meetodite väljatöötamist. Järjestus on vaid esimene samm terviklike andmehulkade nagu geenide tuvastamine, valkude struktuuride, molekulaarsete interaktsioonide ja geeniregulatsiooni mudelite loomiseks. Andmete täiustumine võimaldab seada uusi küsimusi ja leida lahendusi uutele probleemidele. Uute andmestike loomiseks on vaja eksperimentaalsete meetodite ning arvutuslike analüüsimeetodite koostööd. Andmete lisandumine võimaldab järk-järgult aru saada bioloogiliste süsteemide ülesehitusest ning toimimisest (Birney, Clamp, & Hubbard 2002).

Bioloogia kui terviku organiseeritust kirjeldavad erinevatest allikatest pärit bioloogilised andmed. Seega on kõige enam väärt süstemaatiliselt organiseeritud ning omavahel integreeritud erinevatelt bioloogilistelt tasemetelt pärit andmed. Omades suurt hulka toorandmeid valkudest, RNAst, aga ka genoomi järjestustest ja struktuuridest, valgu ja RNA ekspressioonimustritest ning rakulistest asukoha kujutistest, on tekkinud suur vajadus hoida, väärtustada ja tagada ligipääs informatsioonile. Erinevate andmebaaside integreerimine teeb võimalikuks andmete puudujääkide lihtsa identifitseerimise ning seeläbi nende kiirema kõrvaldamise ja andmete täiustamise järjekorra väljatöötamise.

Peamine väljakutse andmebaaside arenduses on andmete ühendamine võimaldamaks info paremat levikut tulenevalt genoomide täielikust sekveneerimisest. Üks peamisi probleeme, mis lahendust vajab, on algsete infoallikate seotuse säilitamine annoteeritud andmete ning allikate vahel. Sageli uute analüüsimeetodite väljatöötamisega algsed andmed vaadatakse üle ning analüüsitakse uuesti. Tihti aga jäävad nende andmete põhjal genereeritud annotatsioonid muutmata, sest tagasiside puudub allikate ja annotatsioonide vahel.

Ideaalsel juhul peaks kõikide andmebaaside andmed olema seotud stabiilse, versioonipõhise identifikaatoriga ja andmete omavahelised seosed salvestatud, et oleks algandmete muutuse järgselt võimalik annotatsiooni uuendada (Birney, Clamp, & Hubbard 2002). Oluline on andmebaaside loomisel sil-

mas pidada ka andmete täielikkust ja kvaliteedi ulatust (Birney, Clamp, & Hubbard 2002). Ideaalsel juhul koosneb andmebaas bioloogiliste eksperimentide teel saadud andmetest ning nende põhjal *in silico* teostatud analüüsi tulemustest. Tegelikkuses on selliseid eksperimentaalselt tõestatud andmetel põhinevaid andmebaase väga vähe. Andmete esitamisel ja kasutamisel on oluline märkida andmete päritolu ning saamisviis, olgu selleks eksperimentaalne või *in silico* ennustus, ja eeldatav täpsus või muu kvaliteedi hinnang ning viimane uuendamise aeg.

Seega on andmebaaside loomise seisukohast peamised etapid:

- andmete kogumine
- andmete töötlemine
- organiseerimine
- hindamine
- seoste loomine erinevate andmete vahel
- olemasoleva info põhjal uute andmete genereerimine

1.2.2 Relatsioonilised andmebaasid

Geneetilise informatsiooni haldamine nõuab pikaajalist andmete säilitamist ja hõlpsalt programmeeritavaid viise informatsiooni uuendamiseks ja ligipääsuks. Enamus bioloogilisi andmebaase kasutavad relatsioonilist andmebaaside haldamise süsteemi (RDBMS) kui põhilist andmete haldamisviisi. RDBMS eelised on järgnevad (Birney, Clamp, & Hubbard 2002):

- kahe viimase aastakümne jooksul on arvutiteaduses loodud hästi arusaadavad, kergesti käsitletavad, viimistletud süsteemid, mis võimaldavad andmete terviklikkust.
- relatsioonilised andmebaasid kasutavad standardiseeritud päringu keelt (SQL) ja kõik põhilised programmeerimiskeeled on seotud SQL kasutajaliidesega, võimaldades programmilist ligipääsu.
- on suur hulk RDBMS-i ja SQL-i valdavaid spetsialiste, keda on võimalik kaasata RDBMS-põhinevatesse andmebaasilahenduste väljatöötamisse. Enamus andmebaase põhinevad kas Oracle, Sybase, Postgres, IBM DB2, mSQL või MySQL-il. Seejuures tuleb märkida, et MySQL ei ole küll rangelt RDBMS kuid praktilistel kaalutlustel võib seda pidada RDBMS-ks ning MySQL on laialdaselt kasutusel bioloogilistes andmebaasides.

1.2.3 Andmebaaside kasutajaliidesed

Enamuse andmebaaside puhul pole tavakasutajale antud otsest võimalust programmiliseks ligipääsuks, selle asemel on mitmed tarkvarakihid, mis on loodud andmebaasi peale. Üldistused, mida kutsutakse programmi rakenduslikuks kasutajaliideseks (API), või vahetarkvara kihiks, võimaldavad andmebaasi skeemi isoleerida süsteemi programmeeritud klientidest. Vahetarkvara kihid erinevad oma keerukuselt ja väljanägemiselt. Mitmed andmebaasid kasutavad BioPerli või BioJava rakendustel põhinevaid kasutajaliidese tuumi. Selline laialdane ühtsete kasutajaliideste kasutamine võimaldab ühendada ja sobitada komponente omavahel ja seega väheneb komponentide ühendamisel tekkivate probleemide hulk. Enamus andmebaaside veebi-põhiseid kasutajaliideseid on dünaamiliselt programmeritud, võimaldamaks andmete kujutamist sõltuvalt konkreetselt päringus soovitud andmetest.

Kui tavakasutajat rahuldab ligipääs andmetele veebi kaudu, siis korralikuks andmeanalüüsiks läheb vaja enamat. Erinevad andmebaasid võimaldavad mitmesugust ligipääsu andmetele, mis varieerub andmebaasi terviklike andmefailide jagamisest tekstifailide (Wingender *et al.* 2000), Exceli tabelite või teiste määratlemata formaatidena.

1.2.4 Geeniregulatsiooni andmebaasid

Geeniregulatsioon hõlmab paljusid kudesid, rakke, arengujärke, keskkonnatingimusi ning kõigi nende kombineerimist ja analüüsimist pole võimalik eksperimentaalselt läbi viia. Seepärast vajatakse tööriistu, mis aitaks analüüsida geeniregulatsiooni mõjutavaid faktoreid ning modelleeerida *in silico* geeniregulatsiooni etappe.

Juba pikka aega on kogutud geeniregulatsiooni kirjeldavaid andmeid andmebaasidesse (Ghosh 1990). Olemasolevad andmebaasid sisaldavad genoomide ja regulatoorsete elementide järjestusi,nende kirjeldusi ning omavahelisi seoseid. Vajalikud on sellised andmebaasid nii biotehnoloogias, farmakoloogias kui mujal teadusharudes. Kõik olemasolevad andmebaasid hõlmavad mingit osa kogu geeniregulatsiooni valdkonnast ning kattuvad omavahel paljudes andmetes, kuid siiski puudub ühtne integreeritud platvorm, mis hõlmaks kõiki olemasolevaid andmeid geeniregulatsiooni kirjeldamiseks ja geenivõrkude modelleerimiseks. Andmebaasid on aluseks geeniregulatsiooni mehhanismide modelleerimiseks, transkriptsioonifaktorite omavaheliste seoste leidmiseks, kirjeldamiseks ning uute *in silico* andmete tootmiseks. Siiani on suurimateks pärmi transkriptsioonifaktoreid hõlmavateks andmebaasideks EPD (Perier *et al.* 2000; Praz *et al.* 2002), SCPD (Zhu & Zhang 1999), SGD (Dwight *et al.* 2002), TRANSFAC (Wingender *et al.* 2000), TRRD (Kolcha-

nov *et al.* 1999; 2000).

Senini olemasolevates *S. cerevisiae* transkriptsioonifaktorite andmebaasides on puudunud võimalus ühtseks päringuks üle kõikide erinevate seostumissaitide tüüpide. Tavaliselt on esitatud eraldi stringid, konsensusjärjestused ning maatriksid (Zhu & Zhang 1999; Wingender *et al.* 2000). Selline esitusviis ei ole aga kasutajasõbralik ning on ebainformatiivne.

Peatükk 2

Meetodid

2.1 *In vitro* transkriptsioonifaktorite seostumissaitide määramine

Rakkudes toimuv geeniekspressiooni jooksev ümberprogrammeerimine, mis on seotud rakutsükli ning keskkonnamuutustega, on jälgitav DNA spetsiifiliste regulaatorite ja DNA vahelise seostumise muutumisega. Erinevad DNA seostumisvalgud on seotud DNA tsentromeeridega, telomeeridega ja teiste regulatoorsete piirkondadega, kus nad reguleerivad kromosoomi replikatsiooni, kondensatsiooni, sidusust ja teisi genoomi säilitamise aspekte.

Viimased edusammud on näidanud kromatiini keskkonna kriitilist rolli geeniekspressiooni regulatsioonis. Histoonivalkude modifitseerimine võib suunata transkriptsioonifaktoreid seonduma spetsiifilistele DNA regioonidele (Jenuwein & Allis 2001). Tuuma histoonivalkude spetsiifiline atsetüleerimine või metüleerimine võib mõjutada transkriptsioonifaktorite seondumist (Weinmann & Farnham 2002). On laialdaselt täheldatud, et hüperatsetüleeritud piirkonnad genoomis on valkude seondumiseks rohkem kättesaadavad kui hüpoatsetüleeritud saidid. Järelikult sama primaarjärjestus võib olla äratuntav transkriptsioonifaktori poolt, kui on hüperatsetüleeritud, ning vastupidiselt võib olla transkriptsioonifaktori poolt äratundmatu, kui genoomijärjestus on hüpoatsetüleeritud (Weinmann & Farnham 2002). Seega on elusate rakkude uurimisel saadud informatsioon transkriptsioonifaktorite seondumistele genoomiga kõige täpsem.

Ekspressiooniandmete analüüs DNA mikrokiibiga võimaldab uurijatel identifitseerida mRNA hulga muutuseid elavas rakus erinevatel tingimustel ning antud meetodi abil on võimalik teada saada, millisel ajahetkel ning millises koes geen ekspresseerub. Nende andmete põhjal on võimalik teha järeldusi geeni funktsiooni kohta. Samas annab geeniekspressiooni uurimine, põhine-

des geenieksрressiooni mustril, võimaluse analüüsida raku seisundit (DeRisi, Iyer, & Brown 1997). Kuigi valkude olemasolu rakus ei sõltu ainult mRNA regulatsioonist on siiski üldistatult võimalik siduda raku tüübi ja seisundi erinevusi mRNA hulga muutusega rakus.

Paljud valgud seonduvad spetsiifilistele saitidele genoomis, et reguleerida geenide ekspressiooni ja säilitamist. DNA spetsiifilised regulaatorvalgud seostuvad spetsiifilisele promooterjärjestusele ja aktiveerivad kromatiinmuutjaid komplekse ja transkriptsiooniaparaati ning initseerivad sellega RNA sünteesi (Ptashne & Gann 1997; Lee & Young 2000; Malik & Roeder 2000)

DNA ja valgu vaheliste seostumissaitide tuvastamiseks on neli peamist meetodit: DNA–valk kompleksi liikuvuse muutus (*retardation*) geelil, DNaas I jalajälg (*DNase I footprinting*, interferentsi analüüside modifitseerimine (*modification interference assays*)ning kromatiini immunosadestamine (*chromatin immunoprecipitation*). Järgnevalt esimese kolme meetodi lühiülevaated (Brown 2001) järgi.

2.1.1 DNA–valk kompleksi liikuvuse muutus (*retardation*) geelil

DNA–valk kompleksi liikuvuse meetodi puhul valk-DNA interaktsiooni moodustumisel kasvab tekkiva kompleksi molekulaarmass ning seda muutust on võimalik identifitseerida elektroforeesil. DNA ülesvoolu järjestus lõigatakse restriktsiooni endonukleaasiga ja seejärel seotakse regulaatorvalguga. Restriktsioonifragment, mis sisaldab kontrolljärjestust moodustab regulaatorvalguga kompleksi, ülejäänud fragmendid jäävad seostumata. Kontrolljärjestuse asukoht määratakse restriktsiooni kaardilt vastavalt fragmentide lahutusele elektroforeesil. Lahutusvõime sõltub restriktsioonikaardi täpsusest ning kui sobivalt on restriktsiooni saidid asetunud. Kahjuks antud meetodi lahutusvõime ei suuda alati määrata kontrolljärjestuse asukohta täpselt ja selle analüüsiks on vaja spetsiifilisemaid meetodeid. Selliseks meetodiks sobib DNaas I jalajälg.

2.1.2 DNaas I jalajälg

DNaas I jalajälje meetod põhineb regulaatorvalgu interaktsioonil DNA-ga, mis kaitseb DNA-d DNaas I endonukleaasse aktiivsuse eest. Meetodi käigus märgistatakse DNA fragment ühest otsast radioaktiivse markeriga. Seejärel seotakse regulaatorvalgud DNA-ga ning lisatakse DNaas I-te piiratud koguses, et tekiksid osalised fragmentide moodustumised. Eesmärk on lõigata iga molekuli vaid üht fosfodiestersidet. Kui DNA fragment ei ole seotud regulaatorvalguga, siis tekivad ühenukleotiidse erinevusega fragmendid. Tekkinud

fragmendid saab eraldada polüakrüülamiidgeelil. Autoradiograafial moodustub vöötide redel. Kui aga regulaatorvalk seostus DNA-ga, siis kaitses ta DNA-d endonukleaasi eest ning fosfodiestersidemed jäid terveks. Puuduvate vöötide järgi saab leida "jalajälje" ehk DNA piirkonna, kuhu regulaatorvalk seostus. Fragmendi suuruse saab välja arvutada "jalajälje" kõrval asuvate vöötide pikkuste järgi. Meetodi peamine puudus seisneb selles, et ei ole võimalik leida milline valk seostus spetsiifilisele järjestusele (Kang, Vieira, & Bungert 2002).

Kaks eelnevat meetodit võimaldavad küll leida seostumisjärjestused, kuid ei anna infot seostuva valgu ja DNA vahelise interaktsiooni kohta. DNaas I annab infot DNA regiooni kohta, mis on seostunud valgu poolt kaitstud. Valgud on aga suhteliselt suured võrreldes DNA kaksikheeliksiga ja seega võivad valgud kaitsta mitmeid kümneid aluspaare, kuigi ise on seostunud DNA-ga vaid mõne aluspaarilisel järjestusel. Seega ei piiritle "jalajälje" meetod täpselt regulaatorpiirkonda, vaid määrab regiooni, milles see asub.

2.1.3 Interferentsi analüüside modifitseerimine

Nukleotiidid, mis moodustavad valguga komplekse, saab määrata modifitseeritud interferentsi analüüsidega. Sarnaselt DNaas I jalajäljele, tuleb DNA fragmendid ühest otsast märkida. Seejärel töödeldakse fragmente kemikaalidega, mis mõjuvad vaid kindlale nukleotiidile. Näiteks dimetüülsulfaat, mis lisab metüülgrupid guaniini nukleotiididele. Selline muutmine toimub piiratud tingimustes, et keskmiselt muudetaks üht nukleotiidi DNA fragmendi kohta. Seejärel DNA segatakse valgu ekstraktiga. Analüüs põhineb sellel, et seostuv valk tõenäoliselt ei seostu DNA-ga kui guaniin on kontrollregioonis muudetud, kuna nukleotiidi metüleerimine segab spetsiifilist keemikalist reaktsiooni, mis võimaldab moodustuda valk-DNA kompleksil. Puuduva valk-DNA seose tuvastamine toimub agaroosi–geelelektroforeesil, kus kaks vööti vastavad DNA-valk kompleksile ning üks ilma valguta DNA-le. Vööt, mis vastab seostumata DNA-le puhastatakse geelilt ja töödeldakse piperidiiniga, mis seob DNA molekulid metüleeritud nukleotiididele. Seejärel saadud produktid lahutatakse polüakrüülamiidgeelil ja tulemused visualiseeritakse autoradiograafiaga. Vöötide suurus viitab DNA fragmendi guaniinindele, mille metüleerimine hoidis ära valgu seostumise. Guaniinid asuvad kontrolljärjestustes. Seejärel muudetud analüüsi võib korrata kemikaalidega, mille sihtmärkideks on A, T või C nukelotiidid ja selle abil piiritleda täpselt regulaatorjärjestus.

Regulaatorjärjestuste olemasolu kontrollitakse ning funktsiooni uuritakse deletsiooni analüüsidega. Meetod põhineb eeldusel, et regulaatorjärjestuse deletsioon viib ekspressiooni muutusele. Kasutatakse reportergeene, mis on viidud kloonitud geeni ülesvoolu järjestusse ning mis asendab olemasoleva

geeni. Kloonitult peaks reportergeeni ekspressiooni profiiil täpselt jäljendama originaalgeeni, kui reporter geen on täpselt sama kontrolljärjestuse mõju alla kui originaalgeen. Reportergeeni valimisel tuleb jälgida, et geeni fenotüüp ei tohi olla juba avaldunud peremees organismis, et fenotüüpi on kerge detekteerida ja kui võimalik, siis oleks fenotüüpi võimalik kvantitatiivselt mõõta. Kloneeritud reportergeeni ülesvoolu järjestustest lõigatakse võimalikke regulaatoralasid. Seejärel viiakse muudetud konstrukt peremeesorganismi ning jälgitaks geeniekspressiooni mustri muutust. Kui geeniekspressioon läheb üles, siis lõigati ära repressor või vaigistaja, kui geeniekspressioon läheb alla, siis lõigati välja aktivaator või võimendaja ning kui muutus koespetsiifilisus, siis see viitab koespetsiifilisele regulaatorjärjestuse eemaldamisele.

Eelnevalt kirjeldatud meetodid sobivad juba piiritletud regulaatorregioonide täpsemaks uurimiseks. Suuremahulist uuringut nende meetoditega ei saa teostada liigse ajamahukuse tõttu. Seega on oluline, et oleks võimalik määrata suuremahulisel uuringul eelnevalt tõenäosuslikud regulaatorpiirkonnad. Peamine meetod, mille abil tänapäeval ülegenoomselt valk-DNA interaktsioone uuritakse on kromatiini immunosadestamine. Järgnevalt käsitletakse seda pikemalt.

2.1.4 Kromatiini immunosadestamine kiibil

Kromatiini immunosadestamine kiibil (*Chromatin immunoprecipitation* (ChIP on Chip)) võimaldab jälgida valk–DNA interaktsioone üle terve pärmi genoomi ning ülegenoomsel asukoha analüüsil võimaldab leida transkriptsioonifaktorite seostumisi *in vivo* ja seeläbi analüüsida regulatoorseid võrke (Qiu 2003; Ren *et al.* 2000). Ülegenoomne seostumiste esinemine ja ekspressiooniandmete kombinatsioon võimaldab identifitseerida üldist geenide hulka, mille ekspressioon on otseselt kontrollitud transkriptsiooni aktivaatorite poolt rakus (Ren *et al.* 2000). Meetod põhineb muudetud kromatiini immunosadestamisel, mida on varem kasutatud uurimaks väikesel hulgal spetsiifilisi DNA saite valk–DNA interaktsioonil, koos DNA mikrokiibi analüüsiga. ChIP on Chip meetodi peamine puudus seisneb selles, et antikehadega rikastatud DNA seostumine regulaatorvalkudega ei viita alati sellele, et valk seostub sadestatud järjestusega vaid pigem viitab valk-valk interaktsioonidele (Kang, Vieira, & Bungert 2002). ChIP on Chip meetod suudab määrata seostumise 1-2 kiloaluspaarilise täpsusega (Liu, Brutlag, & Liu 2002).

Meetodi lühiülevaade

Rakud kinnistatakse formaldehüüdiga, kogutakse ja töödeldakse ultraheliga. DNA fragmente, mis on ristseotud(*cross-linked*) meid huvitava valguga ri-

kastatakse immunosadestamisel spetsiaalse antikehaga. Seejärel rikastatud DNA amplifitseeritakse ja märgistatakse fluorestseeruva värviga (Cy5) kasutades ligeerimisvahendatud polümeraasi ahelreaktsiooni. DNA proov, mida ei rikastatud immuunosadestamisel, värvitakse Cy5-st erineva fluorestseeruva värviga ja allutatakse ligeerimisvahendatud polümeraasi ahelreaktsioonile. Nii immunosadestamisel rikastatud kui ka rikastamata proovid hübridiseeritakse DNA mikrokiibile, mis sisaldab kõiki pärmi intergeenseid järjestusi. Kolmelt eraldisesivalt immuunosadestamise eksperimendilt saadud fluorestsentsmärgiste intensiivsuse tasemed analüüsitakse kaalutud keskmise meetodiga, leidmaks valgu suhtelist seondumist kiibi iga järjestusega. Tugevalt rikastatud järjestused on tavaliselt tõelised sihtmärgid, ning neis esinevad sagedasti transkriptsioonifaktorite seondumissaidid (Weinmann & Farnham 2002).

ChIP on Chip meetodi poolt leitud kandidaatregioonid analüüsitakse motiiviotsimis algoritmidega. Üheks näiteks võib tuua *Motif Discovery scan(MDscan)* algoritmi, millega analüüsitakse kiibilt saadud järjestused ning otsitakse DNA motiive, mis võiksid esitada valk-DNA interaktsioonisaite (Liu, Brutlag, & Liu 2002). MDscan kasutab motiiviotsimisel sõnade loendamist ning positsiooni spetsiifilise kaalumaatriksi uuendamist (Liu, Brutlag, & Liu 2002).

2.2 *In silico* analüüs

Tavaliselt kasutatavad motiivide esitusviisid on positsiooni spetsiifiline skoorimaatriks (PSSM) ning Rahvusvaheline Puhta ja Rakenduskeemia Liidu poolt välja töötatud IUPAC koodi kasutav PROSITE tüüpi regulaaravaldis. PSSM salvestab iga DNA nukleotiidi eelistuse igas seostumissaidi positsioonis. Selline esitus põhineb eeldusel, et positsioonid maatriksis on teineteisest sõltumatud. Puudub ühene seisukoht, kas nii tugev sõltumatuse eeldus on põhjendatud. Viimased tulemused viitavad, et mõningatel juhtudel esineb positsioonide vahel sõltuvus (Barash *et al.* 2003). Vähem väljendusrikkad mudelid ei suuda esindada keerukaid sõltuvusi, kuid neid võib õppida väikese hulga näidete põhjal. Rohkem väljendusrikkad mudelid suudavad esitada keerukamaid sõltuvusi, kuid kaasavad mitmeid parameetreid ning nõuavad suuremat näidete hulka õppimiseks.

Enamik transkriptsioonifaktorite seondumissaite on saadud *in silico* ennustamisel. Peamiselt teostatakse *in silico* ennustusi geenieksprессiooniandmete analüüsil ja fülogeneetilisel jalajäljel põhinevate meetoditega.

2.2.1 Fülogeneetiline jalajälg

Fülogeneetilise jalajälje teooria põhineb erinevate organismide genoomijärjestuste analüüsil. Teooria aluseks on eeldus, et funktsionaalsed osad genoomist muteeruvad valikulise surve all aeglasemalt kui mittefunktsionaalsed järjestused. Genoomide ortoloogsete regulaatorpiirkondade võrdlemisel leitavad konserveerunud järjestused on tavaliselt head kandidaadid funktsionaalsete regulaatoralade tuvastamisele. Fülogeneetilise jalajälje peamine eelis üksiku genoomi geenidel põhineva ennustuse ees on, et puudub vajadus usaldusväärse meetodiga leitavate koreguleeritud geenide hulga järele. Vastupidiselt, fülogeneetilise jalajälje meetodiga on võimalik identifitseerida regulatoorseid elemente isegi üksikule geenile, kui regulaatorelemendid on vajalikul määral konserveerunud üle mitmete liikide (Blanchette & Tompa 2002).

Fülogeneetilise jalajälje koostamine

Standardmeetodina, mida kasutatakse fülogeneetilise jalajälje koostamisel, konstrueeritakse ortoloogsete regulaatorjärjestuste globaalne mitmene joondamine ja seejärel tuvastatakse joonduses konserveerunud järjestused. Antud meetodi probleem seisneb selles, et kuna regulatoorsed alad, mis on 5–20 aluspaari pikad, on väga lühikesed võrreldes regulaatorpiirkondadega, mille pikkusteks loetakse enamasti 1000 aluspaari. Antud järjestuste pikkuste juures, kui liigid on fülogeneetilises puus mõnevõrra lahknenud, on tõenäoline, et lahknenud mittefunktsionaalne taust ületab lühikese konserveerunud signaali. Selle tulemusena ei joondu lühikesed regulatoorsed elemendid kokku (Blanchette & Tompa 2002). Antud juhul regulatoorsed elemendid ei pruugi kuuluda konserveerunud regioonidesse ning jäävad märkamatuks. Seega, kui regulatoorsed alad on hinnatud keskmiselt kuni kõrgelt lahknenuks, siis globaalne mitmene joondus tõenäoliselt ei leia olulisi signaale (Blanchette & Tompa 2002).

Genoomid tuleb valida põhimõtte järgi, et nad ei oleks liiga lähedased ega liiga kauged evolutsioonilises puus. Liiga lähedaste genoomide puhul suudetakse küll järjestused hästi joondada, kuid funktsionaalsed elemendid ei ole märgatavalt paremini konserveerunud ning seega ei saa neid eristada mittefunktsionaalsetest regioonidest (Cliften *et al.* 2001; Blanchette & Tompa 2002). Liiga kaugete genoomide puhul on aga raske või võimatu leida vigadeta joondust (Tompa 2001; Blanchette & Tompa 2002).

2.2.2 Geenieksprессiooni andmete analüüs

Geenieksprессiooni andmete analüüs põhineb DNA kiibi tehnoloogial. Mikrokiibi tehnoloogia mRNA populatsiooni hulga mõõtmiseks rakkudes võimaldab meil jälgida tuhandete geenide ekspressiooni tasemeid üheaegselt. Mõõtes sadadel erinevatel tingimustel või ajamomentidel ekspressiooni tulemusi on võimalik geenieksprессiooni kaardi koostamine (Brazma *et al.* 1998).

Geenieksprессiooni eksperimentaalses osas *in vitro* seotakse klaaskandjale polümeraasi ahelreaktsiooni (PCR) tehnoloogial ülesamplifitseeritud genoomsed DNA järjestused. Rakukultuurist eraldatakse mRNA. cDNA märgistatakse fluorestsentsmärgisega seotud desoksüuridiin-trifosfaadiga ($dUTP$) ja hübridiseeritakse kiibil olevate oligotega. Seejärel mõõdetakse fluorestsentsvärvuse intensiivsust ning saadud andmed salvestatakse *in silico* analüüsiks (DeRisi, Iyer, & Brown 1997).

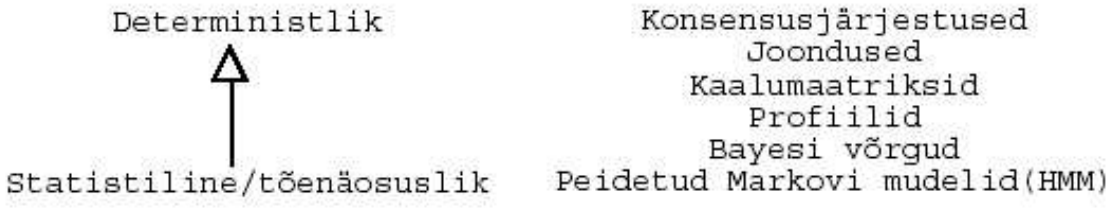
Eksperimentaalselt saadud andmed klasterdatakse, otsitakse järjestuse mustreid geenide ülesvoolu piirkondadest, teostatakse kontrolleksperimendid mustrite olulisuse läve tuvastamiseks, valitakse statistiliselt huvitavad mustrid, saadud mustrid grupeeritakse, esitatakse ühtsel kujul ning leitud tõenäosuslikke mustreid võrreldakse andmebaasis olevate regulatoorsete signaalidega (Vilo *et al.* 2000; van Helden, André, & Collado-Vides 1998; Brazma *et al.* 1998). Geenieksprессiooni *in silico* andmete analüüsi raskuspunkt langeb klasterdamismetoodikale. Valim, millest hakatakse otsima võimalikult suure tähtsusega motiive peab olema optimaalne (Brazma *et al.* 1998).

Geenieksprессiooni andmete analüüsil saadud mustrid esitatakse sageli regulaaravaldisena, sest üksikute oligote esitamisel saame sarnaseid ühe-kahe nukleotiidise erinevusega järjestusi väga palju ning nende kõigi objektiivsuse hindamine on tülikas (Vilo 2002).

2.3 Transkriptsioonisaitide esitamisviisid

Transkriptsioonifaktorite seondumissaite esitatakse mitmel erineval moel. Teistest selgelt paremat esitusviisi pole välja töötatud, igal viisil on omad plussid ja miinused. Otsimisel üle genoomi tuleb välja kaalumaatriksi peamine eelis – väljendusvõimsus võrreldes oligote, konsensusjärjestuste ning regulaaravaldistega. Reaalsel andmestikul on sõltuvusel põhinevad mudelid üldistatult paremad kui PSSM-d. Järgnevalt esitatakse erinevad esitamisviisid ning nende puudused ja eelised.

Erinevad esitusviisid jagunevad ettemääratute (deterministlike) ja statistiliste vahele:



Joonis 2.1: Transkriptsioonisaitide esitusviiside jaotus deterministlike ja statistiliste vahele

2.3.1 Oligonukleotiidid

Oligonukleotiidi ehk stringi tüüpi avaldised pärinevad enamasti eksperimentaalselt tõestatud andmetest. Samuti on stringi kujul esitatud enamik andmebaasides olevaid seondumisaite. Puuduseks on vigadeta otsimine ehk otsimisel leiab ainult 100% sama oligonukleotiidi. Stringide puhul vajab märkimist teisenduskaugus ehk minimaalne arv teisendusi, mida on vaja, et saada stringist A string B. Vähimat arvu insertsioone, deletsioone või asendusi, mida läheb stringi A teisendamiseks stringiks B vaja, nimetatakse muutmiskauguseks ehk Levenshteini kauguseks.

Näidisoligod

Olgu esitatud kuus näidisoligot, mida kasutame ka edaspidi seostumissaitide esitusviiside kirjeldamiseks:

```
TACGCT
TCAGCT
AACGGT
TCCGCA
TCACCT
TCCGGT
```

Levenshteini kaugus näites 2.3.1 toodud viienda (TCACCT) ja kuuenda (TCCGGT) oligo vahel on 3 ühikut.

2.3.2 PROSITE tüüpi regulaaravaldised

Regulaaravaldised, mida kasutatakse bioloogiliste saitide kirjeldamiseks, on vaid osa üldistest regulaaravaldistest. Enamasti bioloogias kasutatakse

PROSITE tüüpi mustreid. Regulaaravaldised, mis kirjeldavad transkriptsioonifaktoreid, sisaldavad järgmisi omadusi:

IUPAC koodi tähised aminohapete ning nukleotiidide kirjeldamiseks

[] kirjeldamaks lubatud aminohappeid, näiteks [TED] puhul on lubatud treoniin, glutamaat või aspartaat.

{} kirjeldamaks mitte lubatud aminohappeid, näiteks {LAP} puhul on lubatud kõik aminohapped peale leutsiini, alaniini ja proliini.

() abil kirjeldatakse lubatud aminohappe kordusi, näiteks P(3) puhul on 3 järjestikkust proliini.

Regulaaravaldise näidis

Peatükis 2.3.1 kirjeldatud näidisoligote põhjal genereeritud regulaaravaldis:

```
[TA][CA](2)[GC](2)[TA]
```

Lahtiseletatult: esimeses positsioonis võib esineda T või A, teises ning kolmandas positsioonis võivad olla C või A nukleotiidid, neljandas ja viiendas positsioonis G või C nukleotiidid ning kuuendas positsioonis kas T või A.

2.3.3 Konsensusjärjestused

Konsensusjärjestuse mõistet kasutatakse laialdaselt esindamaks transkriptsioonifaktorite spetsiifikat. Üldiselt iseloomustab konsensus järjestust, mis sobitab kõik kirjeldatavad saidid peaaegu, aga pole nõutav, et määraks kõiki. Määratakse kompromissiga lubatud mittesobitumised, konsensusjärjestuse mitmesus ja esituse täpsus (Stormo 2000). Konsensusjärjestusega on küll lihtne esitada teatud hulka saite, kuid on keeruline leida konsensusjärjestust, mis oleks optimaalne ennustamaks uute saitide esinemisi. Konsensus väljendab parimat esinemist, mis on arvutatud maatriksi iga positsiooni enamesineva nukleotiidi järgi. Arvutuspõhimõte on, et igas positsioonis võetakse see nukleotiid, mida on esinenud kõige rohkem. Konsensuses esinevad sageli ka lisaks nukleotiidile vastavate tähtede ka muud sümbolid. Selgituseks Rahvusvahelise Puhta- ja Rakenduskeemia liidu poolt väljatöötatud (IUPAC) tabel 2.1 nukleotiidide mitmesuse esitamiseks (Cornish-Bowden 1985):

Tabel 2.1: Laiendatud DNA / RNA tähestik

Sümbol	**Tähendus**	**Aminohape**
A	A	Adeniin
C	C	Tsütosiin
G	G	Guaniin
T	T	Tümiin
U	U	Uratsiil
M	A või C	
R	A või G	
W	A või T	
S	C või G	
Y	C või T	
K	G või T	
V	A või C või G	
H	A või C või T	
D	A või G või T	
B	C või G või T	
X	G või A või T või C	
N	G või A või T või C	

Näidis konsensusjärjestus

Peatükis 2.3.1 kirjeldatud näidisoligote põhjal leitud konsensusjärjestus:

```
TCCGCT
```

2.3.4 Maatriksid

Transkriptsioonifaktorite seostumissaite kirjeldavad maatriksid jagunevad omakorda maatriksiteks, mis väljendavad otseselt nukleotiidide esinemiste arvu, selle meetodi edasiarendusteks ning kaalumaatriksiteks, mis väljendavad erinevate algoritmide abil arvutatud kaale ehk olulisust. Viimane variant võimaldab otsida kindla skoorilävega esinemisi üle genoomi.

Kaalumaatriks (*Position weight matrix*, PWM) on alternatiiv konsensusjärjestusele. Esmalt kasutati kaalumaatrikseid RNA saitide iseloomustamiseks, mis funktsioneerisid *E.coli* translatsiooni initsiatsioonisaitidena (Stormo *et al.* 1982). Leiti, et lisaks Shine-Dalgarno järjestusele ning initsiatsioo-

nikoodonile on ka ribosoomi seostumissaidid kõrgelt konserveerunud (Stormo 2000).

Tabel 2.2: Näite 2.3.1 põhjal koostatud positsioonimaatriks

A	1	2	2	0	0	2
C	0	4	4	2	4	0
G	0	0	0	4	2	0
T	5	0	0	0	0	4

Sellest järeldub, et mitmed aluspaarid ribosoomi seostumisregioonis mRNA-l võivad interakteeruda ribosoomiga ja tõenäosus, et seostumine on piisav initseerimaks translatsiooni, oli kõikide koos toimivate interaktsioonide summa. Saidid, mille kogu koostoime ületas mingi läve, võis lugeda autentseks (*bona fide*) translatsiooni initsiatsioonisaidiks vastupidiselt neile, mis jäid lävest allapoole. Seega sündis kaalumaatriksi idee, esindamaks hulka funktsionaalseid saite ja nendele seostuva valgu spetsiifilisust (Stormo 2000).

Negatiivsete logaritmide meetod maatriksi kaalude leidmiseks

Antud meetod leiab negatiivsed logaritmide abil kaalu iga aluse sageduse kohta igas positsioonis. Konkreetse saidi summa on negatiivne logaritm tõenäosusest, mis väljendab kindla järjestuse esinemisest teatud saitide hulgas eeldusel, et positsioonid on sõltumatud (Staden 1989).

Samuti on näidatud, et on tugev korrelatsioon järjestuse skoori ja promooteri aktiivsuse vahel. Kui kaalud tõesti väljendavad seostumisprotsessi tunnuseid, siis enamate "heade"tunnuste olemasolu peaks viitama kõrgemale aktiivsusele (Mulligan *et al.* 1984).

Kui on olemas piisavalt kvantitatiivseid andmeid järjestuste näol ning nende funktsionaalseid aktiivsuseid, siis peaks olema kergelt lahendatav kaalumaatriksi loomine, mis annaks parima sobivuse sellele kvantitatiivsele andmestikule. Alati ei pruugi parim sobivus olla piisavalt hea. Juhul, kui standardse kaalumaatriksi puhul iga positsiooni skoorid liidetakse, et saada üldine skoor, siis sellest tuleneb, et iga positsioon annab sõltumatu panuse aktiivsusesse. Kui see eeldus on väär, võib isegi parim sobivus anda ebaõige lahenduse. Sellisel juhul on vaja komplitseeritumaid mudeleid, kus maatriksi elemendid vastavad näiteks kahele nukleotiidile mitte ühele. Selline meetod ei leia mitte ainult parimat maatriksit olemasolevale andmestikule, vaid viitab ka seostumise mehhanismile, kus nukleotiidide positsioonid ei ole omavahel sõltumatud. Limiteerivaks on kvantitatiivse andmestiku saamise töömahukus ning seepärast kasutatakse sellist lähenemist väga harva (Stormo 2000).

Informatsiooni sisalduse maatriksid

Olulisel kohal kaalumaatriksite kirjeldamisel on ka informatsioonisisalduse järgi loodud maatriksid. Erinevate regulatoorsete süsteemide seostumissaite võrdlemisel, on välja töötatud välja informatsiooni sisalduse ning selle sõltuvus seostumissaitide sagedusest genoomis (Schneider *et al.* 1986). Informatsiooni sisaldust saidi igal positsioonil võib esitada nii:

$$I_i = 2 + \sum_{b=A}^{T} f_{b,i}\, log_2 f_{b,i} \tag{2.1}$$

kus i on positsioon saidis b viitab võimalikele alustele, $f(_{b,i})$ on iga nukleotiidi leitud sagedus positsioonil i. I_i väärtus on 0 kui kõikide aluste esinemise tõenäosus on 25% ja 2 bitti juhul kui positsioon on täielikult konserveerunud ehk antud positsioonis esineb vaid üks nukleotiid neljast.

Veidi hiljem on näidatud, kasutades statistilise mehhaanika teooriat, et aluste sageduste logaritmid peaksid olema proportsionaalsed nende aluste seostumisenergia panusega (Berg & von Hippel 1987). See teooria toetab informatsiooni sisalduse analüüsi ja soovitab, et informatsiooni sisaldus on seotud saitide hulga keskmise seostumisenergiaga. Pärmi puhul esimene valem viitab positiivsele informatsiooni sisaldusele ja seega spetsiifilisele seostumisenergiale igal juhuslikul saitide hulgal. Parandatud valem, mis võtab arvesse ka pärmis valitsevat nukleotiidide suhet, on järgmine:

$$I_{seq(i)} = \sum_{b} f_{b,i}\, log_2 \frac{f_{b,i}}{p_b} \tag{2.2}$$

kus p_b on aluse b sagedus kogu genoomis. Valem 2.1 on valemi 2.2 erijuht, kus p_b on kõikide b jaoks 0.25. I_{seq} on tuntud kui suhteline entroopia ja Kullback-Liebler kaugus.

Tabel 2.3: Informatsiooni sisalduse maatriks näite 2.3.1 põhjal

A	-2.2	-1.78	-1.78	-2.8	-2.8	-1.78
C	-2.8	-1.18	-1.18	-1.78	-1.18	-2.8
G	-2.8	-2.8	-2.8	-1.18	-1.78	-2.8
T	-0.96	-2.8	-2.8	-2.8	-2.8	-1.18

Valem 2.3 näitab, kuidas leida parimat maatriksit, kui on teada kõrge afiinsusega saidid, kuid pole teada täpset seostumisafiinsust. Eeldame, et teatakse ka organismi täielikku genoomijärjestust, kust valk ja saidid on pärit. Lisamiseeldusest lähtudes, et iga positsioon panustab sõltumatult kogu

seostumisenergiasse, on meil maatriks $H(b,i)$, mis sisaldab seostumisenergia panuseid oma elementidena. Iga üksiku järjestuse S_α kogu seostumisenergia on antud $H(b,i) \cdot S_\alpha$ poolt. Tõenäosus, et valk seonduks saidile järjestusega S_α, arvestades kõiki võimalikke seondumissaite kogu genoomis, on kirjeldatud valemiga:

$$P(S_\alpha \; on \; seostunud) = \frac{e^{-H(b,i)} \cdot S_\alpha}{Z} \tag{2.3}$$

kus Z on alamfunktsioon, 1 kõigi genoomi saitide seostumisafiinsuste summat. Teades, et meie saidid on kõrge seostumistõenäosusega, on järgmine loogiline samm maatriksi leidmine, mis maksimeerib kõikidele saitidele seondumise tõenäosust. Kuna eeldame, et genoom on põhiolemuselt juhuslik, siis me saame arvutada Z-i analüütiliselt (Heumann, Lapedes, & Stormo 1994). Genoomid ei ole juhuslikud järjestused, kuid eeldus on kehtiv kui lühikesed alamjärjestused, seostumissaitide pikkustega, esinevad genoomi aluste eeldatava sagedusega. Sellisel juhul juhuslikkuse eeldus on kehtiv. Antud eeldusele toetudes võib näidata, et maatriksi $H(b,i)$ elemendid, mis maksimeerivad seostumise tõenäosust hulgale funktsionaalsetele saitidele on lihtsad (Heumann, Lapedes, & Stormo 1994).

$$H(b,i) = -ln\frac{f_{b,i}}{p_b} \tag{2.4}$$

Seega, kui on hulk kindla faktori tuntud seostumissaite, siis $-ln\frac{f_{b,i}}{p_b}$ on maksimaalne tõenäosuse hinnang seostumisenergia panusele iga aluse kohta igas positsioonis ja I_{seq} on kõigi tuntud saitide keskmine seostumisenergia (Stormo & Fields 1998).

2.3.5 Bayesi võrgud

Bayesi võrgud on suunatud tsüklivabad graafid, mille tipud esitavad juhuslikke muutujaid ja kaared tõenäosuslikke sõltuvusi tippude vahel (Charniak 1991).

Bayesi võrkude puhul kirjeldatakse iga positsiooni sõltuvust teistest positsioonidest. Näiteks nukleotiidi muutus esimeses positsioonis võib esile kutsuda aminohappe kõrvalahela konformatsiooni muutuse. See aga omakorda võib muuta teiste aminohapete konformatsiooni seostumissaidis ja tingida seostumiseelistuste muutust. Bayesi võrkudega kajastatakse põhjuslikke seoseid orienteeritud graafina ning hiljem analüüsitakse neid. Bayesi võrkude modulaarne süsteem võimaldab kirjeldada lihtsaid eelteadmisi ning erinevaid tõenäosuslikke mudeleid. Samuti on treenitavad Bayesi võrgud võimelised näidetest õppima. Bayesi võrk kirjeldab alati tõenäosusjaotust ning neid

saab genereerida ka väheste näidete põhjal, sealjuures siiski jäädes piisavalt väljendusrikkaks ja kirjeldatuks mitte liiga paljude parameetritega (Barash *et al.* 2003).

2.3.6 Peidetud Markovi mudelid

HMM kirjeldavad süsteemi, mis koosneb eraldiseisvatest olekutest ja olekute vahelistest seostest. Iga seost iseloomustab tõenäosus. Mudelid on "peidetud", kuna seisundeid ei saa otseselt jälgida. Bioinformaatikas on HMM olulised seetõttu, et võimaldavad otsida või luua joonduse algoritmi kindla tõenäosuse baasil ja mudelit on lihtne treenida tuntud andmestikuga (Zhang 2002). Peidetud Markovi mudelid arvestavad tõenäosuse arvutamisel ka eelnevas positsioonis oleva nukleotiidi väärtust ja seega on statistiliselt väljendusrikkamad kui positsioonimaatriksid.

2.4 Andmebaaside modelleerimine

Andmete hoidmiseks, muutmiseks, töötlemiseks ning avaldamiseks läheb vaja seotud andmete kogusid ehk andmebaase. Andmebaaside juhtimissüsteem ehk andmebaasisüsteem (DBMS) võimaldab andmebaasi käsitseda. Andmebaasidega seotud põhitegevused on: andmete hoidmine, lisamine, eemaldamine, parandamine, pärimine. Andmebaasisüsteem peab tagama andmete turvalisuse, terviklikkuse, sünkroniseerumise, andmete taastatavuse ning vältima andmete dubleerumist.

Andmebaaside modelleerimine on vajalik kirjeldamaks olemasolevaid andmehulki, nendevahelisi seoseid ning andmete haldamisega tekkivaid probleeme ja võimalikke lahendusi. Samuti on oluline läbi töötada andmebaasi päringud modelleerimise käigus, võimaldamaks hiljem kõige kiiremaid ning lihtsamaid päringuid. Juba käigusolevate andmebaaside ümbermodelleerimine on tülikas, seega tuleb suurt rõhku panna korralikult toimiva andmemudeli väljatöötamisele.

Peamised modelleerimisel kasutatavad mudelitüübid on olem-seos (*Entity-Relationship*, ER) mudel ning objektmudelid. ER mudeli eesmärgiks on andmebaasi kontseptuaalne kirjeldamine. Relatsioonilise andmemudeli puhul on keskseks objektide väärtusorientatsioon. Erinevate mudelite üheks eesmärgiks on vältida andmeliiasuse tekkimist, kus üht ja sama infot hoitakse erinevates olemites mitmeid kordi. Andmeliiasuse tekkimise vältimiseks on relatsioonilise andmebaasi puhul normaalvormide teooria, mis hõlbustab minimaalse andmeliiasusega skeemi konstrueerimist.

2.4.1 Relatsiooniline mudel

Relatsioonilise mudeli eesmärgiks on seletada andmebaasi põhiolemus lihtsalt ja arusaadavalt, esitada andmete vahelisi seoseid füüsilisest esitusest sõltumatuna ja võimaldada kõrgtaseme andmete manipuleerimiskeeli ehk tehteid relatsioonide kui hulkadega. Oluline on ka andmekaitse ning päringute optimiseerimise võimalus. Samas tekitab relatsioonide paljusus andmete semantikas kadusid ning on oluline probleem tänapäeva kõrge integratsiooniastmega andmete puhul.

Relatsiooniline andmebaas koosneb tabelitest. Iga tabel vastab mingile olemite klassile ning iga tabeli kirje vastab ühele klassi kuuluvale objektile. Iga kirje iga väli kirjeldab klassi kuuluva objekti üht tunnust.

2.4.2 Võtmed, välisvõtmed

Võti on atribuut ehk omadus või ka atribuutide kogum, mis üheselt määrab ära konkreetse olemi. Ühel olemil võib olla mitu võtit kuid enamasti määratakse üks võtmete seast primaarvõtmeks. Ülejäänud võtmed on kandidaatvõtmed. Supervõti on atribuutide hulk, mille pärisalamhulk[1] ei ole võti. Olemi kogumid, millel puudub võti, nimetatakse nõrkadeks olemikogudeks. Tugev olemikogu omab primaarvõtit.

2.4.3 Olem-seos mudel

Olem-seos (ER) mudel on vajalik reaalse maailma kirjeldamiseks ning olemite ja seoste määratlemiseks, enne kui asutakse modelleerima andmebaasi. Olemid on esitatud klassidena, mis kirjeldavad sarnase tunnuse järgi liigitatud olemeid. Olemit iseloomustavad atribuudid ehk tunnused. Seosed ühendavad olemeid omavahel. On olemas kolme tüüpi seoseid:

1:1 seos, kus ühele olemile ühest tüübist vastab üks olem teisest tüübist

1:n seos, kus ühele olemile ühest tüübist vastab n olemit teisest tüübist

m:n seos, kus m olemit ühest tüübist on seotud n olemiga teisest tüübist

Need kvalitatiivsed seosed võimaldavad mudeli seisundi õigsuse kontrolli. Kvalitatiivne tunnus on määratletud modelleerimise käigus ja peab kujutama endast reaalse maailma objektide omavaheliste seoste omadusi.

Atribuutide pärimine toimub ER mudelis üldisemalt olemilt spetsiifilisemale, näiteks geenilt transkriptsioonifaktorile.

[1] Hulga A suvaline alamhulk, mis ei võrdu hulgaga A

Olemid jagunevad veel ka domineerivaks ja alluvaks olemiks. Domineeriva olemi kustutamisel kustutatakse ka alluv olem.

2.4.4 Andmebaasi süsteemi funktsionaalsed komponendid

- Andmete defineerimiskeel (DDL) –kasutatakse andmebaasi struktuuri kirjeldamiseks. Siia kuuluvad:

  CREATE lause ehk tabelite loomine

  ALTER lause ehk tabelite muutmine

  DROP lause ehk tabelite kustutamine

- Andmete manipuleerimiskeel (DML) –kasutatakse andmebaasi protsesside kirjeldamiseks. Siia kuuluvad:

  INSERT lause ehk kirjete lisamine

  UPDATE lause ehk kirjete muutmine

  DELETE lause ehk kirjete kustutamine

  COMMIT, ROLLBACK, SAVEPOINT ehk transaktsioonid andmebaasis

- Andmete pärmiskeel(DQL) –kasutatakse andmete pärimiseks andmebaasist.

  SELECT lause ehk kirjete pärimine

2.4.5 Transaktsioonid ja operatsioonide terviklikkus

Transaktsiooniks loetakse vähimat terviklike omavahel seotud sammude jada, mis võimaldab andmeid muuta. Transaktsioonid on ühtse loogilise terviku moodustavate andmete modifitseerimis (DML)-lausete hulk. Samuti kuuluvad transaktsioonide hulka andmetedefineerimis (DDL)- ja pärimis (DQL)-laused. Transaktsioonide peamised omadused on:

atomaarsus - täidetakse kas kogu transaktsioon või mitte midagi.

isolatsioon - transaktsiooni tulemus peab olema sama, sõltumata kas samal ajal mingeid teisi transaktsioone täidetakse või mitte.

kestvus - kui transaktsioon on lõpetatud, siis ta ei tohi enam kaduma minna.

kooskõla - pärast transaktsiooni lõpetamist peavad andmed jääma samamoodi kooskõlla kui nad olid enne transaktsiooni alustamist.

Transaktsiooni alustatakse esimese SQL lause täitmisel ning lõpetatakse:

- COMMIT või ROLLBACK käsu täitmisel
- DDL või DQL lause täitmisel
- Kasutaja väljalogimisel
- Süsteemi tõrkumisel

Osa II

Praktiline osa

Peatükk 3

Tulemuste arutelu ja analüüs

3.1 Ülesande püstitus

Geeniregulatsiooni mehhanismide mõistmiseks on vajalik transkriptsioonifaktorite ning DNA interaktsioonide kohta infot. Tänapäeval saadakse geeniregulatsiooni andmeid peamiselt kahel viisil: *in silico* ning *in vitro* eksperimentidest. Saadud andmed tuleb töödelda ning esitada parimat modelleerimist võimaldaval kujul.

Olles esmalt õppinud tundma bioloogilisi andmeid kui olemeid ning andmete omavahelisi seoseid elusas looduses, tuleb võimalikult hästi püüda edasi anda neid seoseid ning olemeid ka andmebaasis. Oluline on geeniregulatsiooni andmebaasis võimalikult hästi kirjeldada järgmisi bioloogilisi olemeid:

- transkriptsioonifaktor on valk, mis seondub otse DNA–le geeni *cis*- või *trans*-piirkonnas või reguleerib valk–valk interaktsioonide kaudu geeni ekspressiooni.
- seondumissait on koht kuhu transkriptsioonifaktor seondub DNA–le ning seda seondumissaiti on võimalik esitada nukleotiidide järjestusena
- motiiv on kogum, mis esitab ühe transkriptsioonifaktori seondumissaitide erinevaid nukleotiidseid järjestusi ühtse tervikuna
- geeniregulatsiooni võrgustik on omavahel seotud reaktsioonide võrgustik, mille moodustavad geenid ning nende ekspressiooni mõjutavad transkriptsioonifaktorid

Selliste bioloogiliste seoste esitamiseks lõime **BiGeR**[1] andmebaasi, mis haldab endas geeniregulatsiooni kirjeldavaid andmeid geenide, transkriptsioonifaktorite ja nende seostumissaitide kujul. Andmebaas ei ole mitte ainult

[1] Bioinformatics of Gene Regulation

juba olemasolevate andmete hoidmiseks ning pärimiseks vaid peamiselt edaspidiste *in silico* eksperimentide toetuseks.

Töö peamised etapid

Käesoleva uurimistöö peamised etapid on olnud:

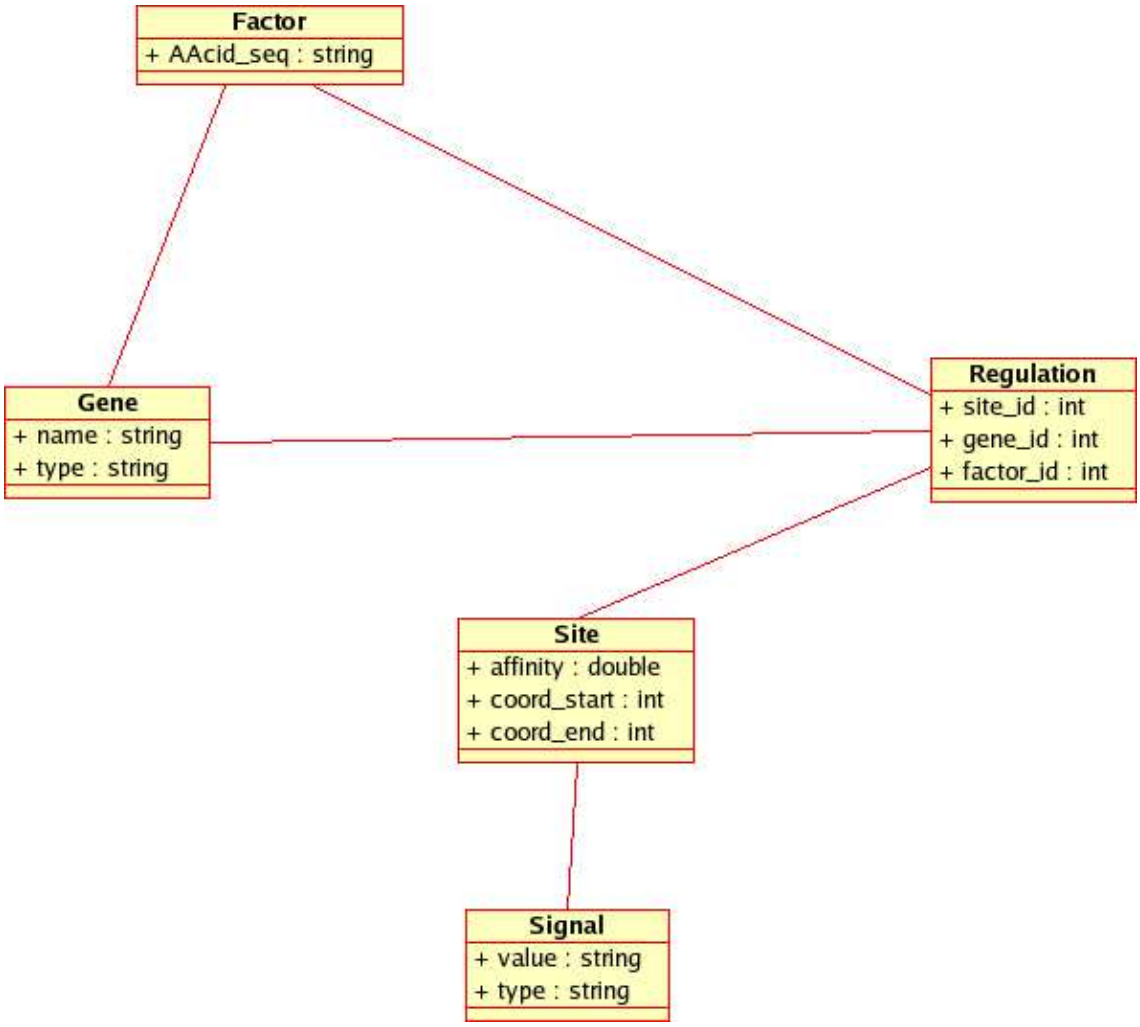
- andmete kogumine
- andmete töötlemine ühtsele kujule
- andmebaasi struktuuri väljatöötamine
- andmebaasi programmeerimine
- andmete automaatne sisestamine andmebaasi

Järgnevas peatükis antakse ülevaade **BiGeR** andmebaasist kirjeldades ära andmebaasi struktuuri tabelite ning atribuutide kujul. Andmebaasi struktuuri hõlpsamaks mõistmiseks on lisatud skemaatilised joonised tabelitest. Andmebaasi funktsionaalsusest antakse ülevaade kirjeldades ära võimalikud kasutusjuhud.

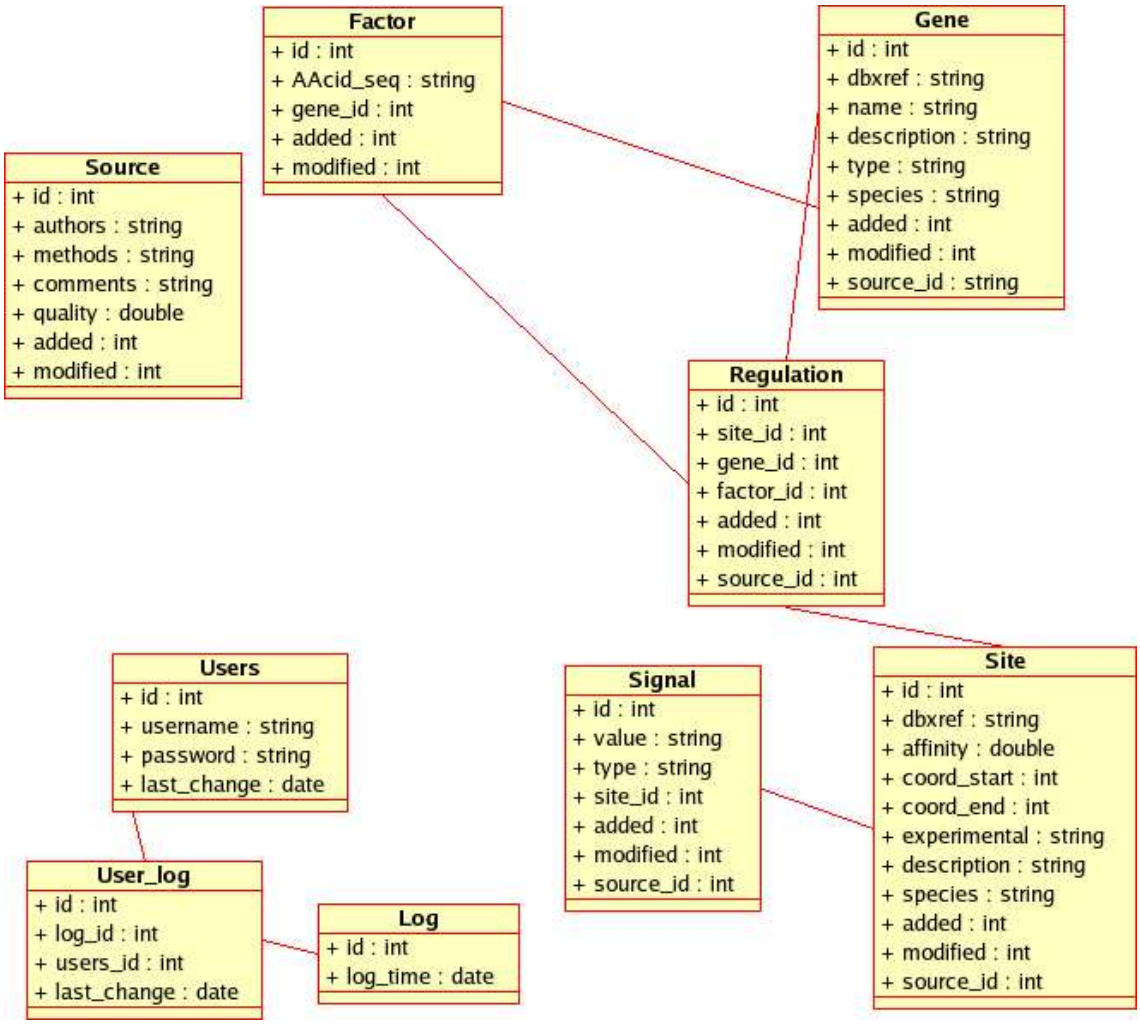
3.2 Andmebaasi skeem

Andmebaasi struktuuri on käesolevas peatükis kirjeldatud kolme joonise abil: joonis 3.1 annab ülevaate **BiGeR** andmebaasi peamistest olemitest ja nende kõige olulisematest atribuutidest, joonis 3.2 kirjeldab ülevaatlikult kõiki andmebaasi olemeid koos kõigi atribuutidega, lisaks on joonisel 3.3 välja toodud kasutajate haldamiseks vajalike tabelite skeem.

Andmebaasi lihtsustatud mudeli joonisel 3.1 on toodud **BiGeR** andmebaasi olulisemad viis olemit ning nende kõige olulisemad atribuudid. Kuna andmebaas on loodud geeniregulatsiooni uurimiseks ja modelleerimiseks ning uute transkriptsioonifaktorite seondumissaitide esitamiseks, on kesksel kohal **Gene, Regulation, Site** ning **Signal** tabelid. Transkriptsioonifaktorite seondumissaitide esitamiseks on kesksed **Site**, **Signal** ning **Regulation** tabelid. **Gene** ning **Factor** tabelis kirjeldatakse ära geenide, nii üldiste kui transkriptsioonifaktoreid kodeerivate geenide, omadused. Geeniregulatsiooni modelleerimiseks peamine tabel on **Regulation**, mis haldab geenide ja transkriptsioonifaktorite vahelisi seoseid, samuti ühendab transkriptsioonifaktorite nende seondumissaitidega. Selline ühtse tabeli kujul geeniregulatsiooni modelleerimiseks vajaminevate andmete esitamine on autorile teadaolevalt senini olemasolevates andmebaasides puudunud.



Joonis 3.1: **BiGeR** peamised tabelid oluliseimate argumentidega



Joonis 3.2: **BiGeR** objekt mudel

Lisaks joonisel 3.1 toodud lihtsustatud mudeli viiele tabelile on olulisel kohal täieliku mudeli joonisel 3.2 toodud **Source** tabel, mis on peamiseks aluseks andmete kvaliteedi hindamise süsteemi väljatöötamisel. Tabelid on omavahel seotud identifikaatoratribuutidega ehk näiteks tabelid **Site** ning **Signal** on seotud **Site** tabeli **id** atribuudi kaudu, kus **Site** tabeli **id** atribuudi väärtus on võrdne **Signal** tabeli **site_id** atribuudi väärtusega. Samuti on oluline kasutajate haldamise süsteem, mis baseerub kolmel tabelil: **User, Log** ja **User_log**.

3.3 Andmebaasi klasside kirjeldused

Käesolevas andmemudelis on kirjeldatud üheksa klassi. Viis neist hoiavad bioloogilisi andmeid, üks kirjeldab andmete allikaid ning kolm tabelit on kasutajate identifitseerimiseks ning andmete lisamis- ja modifitseerimisaegade kirjeldamiseks. Kõikidel klassidel on ühised **id**, **source_id**, **modified** ning **added** väljad.

3.3.1 Tabelite ühised atribuudid

Kõikide tabelite identifikaatoriks ning peavõtmeks on atribuut **id**, mis võimaldab siduda erinevate tabelite andmeid omavahel ning üheselt leida tabeli siseselt kirjeid. Kõikides tabelites, välja arvatud kasutajate haldamiseks mõeldud tabelites, on ühisteks atribuutideks **source_id** mis vastab **Source** tabeli identifikaatorile, **added** ning **modified** atribuudid, mille abil seotakse andmete lisaja(muutja) ning lisamisaeg(muutmisaeg) tabelist **User_log** muudes tabelites olevate andmetega.

3.3.2 Tabel Gene

Gene kirjeldab geenide ehk valku kodeerivate DNA järjestuste lihtsamaid omadusi.

Atribuudid

dbxref on ristviitamiseks vajalik accession number, mis vastab andmete allika identifikaatorile.

name on geeni nimi, tavaliselt suurtäheline lühend.

description sisaldab algallikast pärinevat geeni lühikirjeldust, näiteks artikleid, kust info pärit.

species sisaldab liigi nime, mille andmed on tabelis kirjeldatud.

type on loend (*ENUM*) tüüpi atribuut, mis määrab kas tegemist on geeni(**G**) või faktoriga(**F**).

3.3.3 Tabel Factor

Factor on klassi **Gene** erind. Faktorit eristab geenist eraldi väljatoodud aminohapete järjestus, millelt valk on kodeeritud.

Atribuudid

AAcid_seq on aminohapete järjestus stringi kujul.

3.3.4 Tabel Site

Klass Site kirjeldab transkriptsioonifaktori seondumissaite DNA–l ning transkriptsiooni algussaite (TSS).

Atribuudid

dbxref on viide (*accession number*, AC) andmete allikale ning on oluline interaktsioonide hoidmiseks andmete allika ning käesoleva andmebaasi vahel.

affinity väljendab transkriptsioonifaktori ja DNA vahelise seondumise tugevust. Väärtused on reaalarvulised ja pärinevad ainult eksperimentaalsetest andmetest.

coord_start on seondumissaidi alguskoordinaat DNA–l, alates ORF–i algusest. Välja tüüp on täisarvuline, enamasti negatiivne, väärtus. Negatiivsed väärtused tähistavad ülesvoolu esinemist.

coord_end on seondumissaidi lõppkoordinaat DNA–l, alates ORF–i algusest. Välja tüüp on täisarvuline, enamasti negatiivne, väärtus.

description on andmete üldiseks kirjeldamiseks. Sisaldab infot artiklite kohta, kus antud seondumissait on kirjeldatud.

experimental määrab andmete eksperimentaalse või ennustusliku päritolu. Väli on loend (*ENUM*) tüüpi atribuut, mille väärtused võivad olla **'true'** või **'false'**. Vaikimisi on väärtus **'false'**.

species väärtus on liigi nimi, mille andmed on tabelis kirjeldatud.

3.3.5 Tabel Signal

Signal kirjeldab bioloogiliselt oluliste saitide esinemisi mitmel erineval kujul: näiteks oligonukleotiidid, regulaaravaldised, maatriksid, TSS–d.

Atribuudid

value kirjeldab seondumissaiti, mis on esitatud tabelis **Site**. Välja väärtuseks on vaba tekst,

type kirjeldab signaali esitustüüpi. Väärtuseks võivad olla: **oligo, regular expression, consensus, matrix, TSS**.

3.3.6 Tabel Regulation

Antud klass sisaldab infot **Site** tabelis oleva info seotusest **Gene** tabelis oleva infoga ehk millised seondumissaidid on konkreetsel faktoril või milliste geenide ees antud saidid esinevad. Samuti kirjeldatakse geeniregulatsiooni kujul: faktor A reguleerib geeni B.

Atribuudid

site_id on seoses esineva saidi identifikaator

gene_id on seoses esineva geeni (faktori) identifikaator

factor_id on mõjutava faktori identifikaator, mis pärineb tabelist **Gene**

3.3.7 Tabel Source

Source on klass andmete päritolu kirjeldamiseks ning on aluseks andmete kvaliteedihindamise väljatöötamisel.

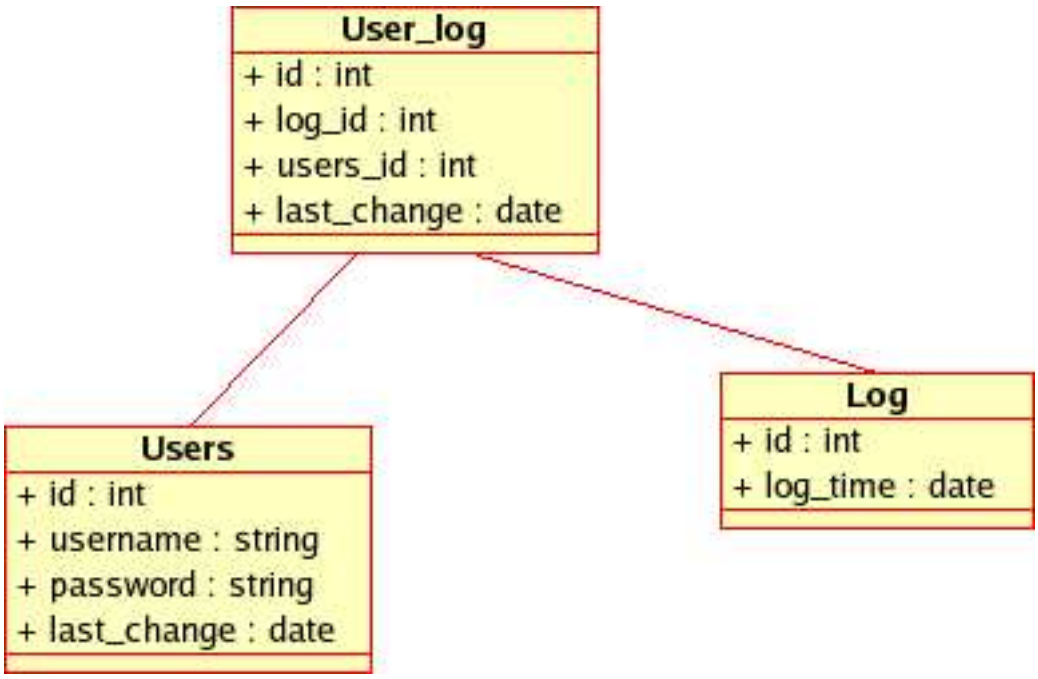
Atribuudid

authors sisaldab andmete autorit identifitseerivat kirjet. Välja tüübiks on vaba tekst.

methods kirjeldab meetodeid, millega andmed on saadud. Välja tüübiks on vaba tekst.

comments väljas võib kirjeldada artikli pealkirja ning ilmumisaja, kust andmed pärit. Väärtuseks on vaba tekst.

quality väljendab kvaliteeti ning on esitatud reaalarvulisel kujul.



Joonis 3.3: Kasutajate autentimiseks ning andmete lisamis- ja muutmisaja haldamiseks vajalikud tabelid

Joonisel 3.3 on toodud kasutajate ning andmete lisamis- ja muutmisaegade haldamiseks vajalikud tabelid. Tabeli **User_log** identifikaatorit kasutatakse teiste tabelite added ning modified atribuutide väärtusena. Tabeli **User** atribuudid **username** ja **password** on vajalikud kasutajate tuvastamiseks.

3.3.8 Tabel Log

Log klass on mõeldud kasutajate logimise haldamiseks.

Atribuudid

Log_time, mis salvestab sisse logimise aja

3.3.9 Tabel User

On eeskätt kasutajate tuvastamiseks ning andmete lisajate kirjeldamiseks loodud klass.

Atribuudid

username hoiab kasutajanimesid

password on vajalik kasutajate üheseks turvaliseks tuvastamiseks

3.3.10 Tabel User_log

Antud klass võimaldab siduda logimisajad kasutajanimedega ja seeläbi on üheselt leitav nii andmete lisaja kui lisamisaeg.

Atribuudid

log_id on klassi Log identifikaator

user_id on klassi User identifikaator

3.4 Kasutusjuhud

Kasutusjuht (*Use Case*) on järjekord toimingutest ja seostest kirjeldatava süsteemi ning selle kasutaja (*Actor*) vahel. Kasutatakse peamiselt süsteemi funktsionaalsete võimaluste väljendamise vahendina. Kogu süsteemi funktsionaalsus määratletakse kasutusjuhtude komplektiga, kus iga kasutusjuht esindab spetsiifilist sündmuste voogu. Kasutusjuhtu võib defineerida ka süsteemi käitumise tegevuse järjestusena, mis annab iga kasutaja puhul jälgitava tulemuse. Seejuures on kasutaja süsteemiväline isik või isend, mis suhtleb süsteemiga vastastikuselt (Cockburn).

Järgnevalt esitatakse bioloogilisi kasutusjuhte, mille abil on võimalik seletatakse **BiGeR** andmebaasi funktsionaalust ning antakse ülevaade andmebaasis realiseeritud päringutega.

3.4.1 Konkreetne transkriptsioonisait kindla geeni ees

Kasutajal on andmed transkriptsioonifaktori kohta ning selle esinemissait konkreetse geeni ees. Eksisteerivad seosed transkriptsioonifaktori ja geeni vahel, samuti transkriptsiooni faktori ja seondumissaidi vahel ning geeni ja seondumissaidi vahel. Vaja on kirjeldada kolm seost:

- transkriptsioonifaktor seondub DNA-le seondumissaidis
- seondumissait on geeni ülesvoolu järjestuses
- transkriptsioonifaktor reguleerib geeni ekspressiooni

Andmete lisamine

Esmalt lisatakse tabelisse **Source** andmete autorit ja saamismeetodit kirjeldavad andmed. Saadud **id** väärtus lisatakse järgnevatesse tabelitesse, atribuudi **source_id** väärtuseks. Teiseks lisatakse transkriptsioonifaktorit kirjeldavad andmed tabelitesse **Gene** ning **Factor**. Samuti kirjeldatakse ära geeni omadused tabeli **Gene** abil. Lisatakse tabelisse **Regulation** transkriptsioonifaktori ja geeni id–d. Seejärel kirjeldatakse vastavalt tabelile **Site** ära seostumissaidi omadused, lisatakse kindlasti väärtused väljadesse: **coord_start, coord_end, experimental, species** ning soovitavalt ka **description**. Viimasena lisatakse andmed tabelisse **Signal**. **Value** atribuut saab väärtuseks DNA järjestuse, **type** on antud juhul **oligo**.

Andmete päring

Olgu soov pärida kõiki regulatsioone, milles osaleb geen GAL4. Andmete päring tuleb lahendada kahes osas: esiteks pärida **Gene** tabelist geeni nimele vastav identifikaator ning seejärel sellele identifikaatorile vastavad regulatsioonid tabelist **Regulation**. Näidispäringud:

```
Päring a:
SELECT  name, id
FROM  Gene
WHERE  name='GAL4'
AND type='G';

Tulemus a:
+------+----+
| name | id |
+------+----+
| GAL4 | 55 |
+------+----+

Päring b:
SELECT  id, site_id, gene_id, factor_id
FROM Regulation
WHERE gene_id='55';
```

```
Tulemus b:
+------+---------+---------+-----------+
| id   | site_id | gene_id | factor_id |
+------+---------+---------+-----------+
| 1010 |    1082 |      55 |       263 |
| 1011 |    1083 |      55 |       263 |
+------+---------+---------+-----------+
```

3.4.2 Transkriptsioonifaktori konserveerunud sekvents ja loetelu geenidest, mille järgi genereeritud

Kasutajal on geeniekspressiooni analüüsi andmete põhjal loodud konsensusjärjestus ning geenide nimekiri, mille järgi konsensusjärjestus genereeritud. Eksisteerivad seosed:

- seondumissait on geeni ülesvoolu järjestuses

Andmete lisamine

Esmalt lisatakse **Source** tabelisse andmete autor, saamismeetod. Saadud **id** väärtus lisatakse järgnevatesse tabelitesse, atribuudi **source_id** väärtuseks. Teisena lisatakse seostumissaiti kirjeldavad andmed, koordinaadid, kirjeldused, liik, afiinsus (kui on väärtus), tabelisse **Site**. Kolmandaks lisatakse konserveerunud sekventsjärjestus tabelisse **Signal**, atribuudi **value** väärtuseks, sealjuures määratakse **type** atribuudi väärtuseks **consensus**. Neljanda etapina lisatakse **Regulation** tabelisse **site_id** ning **gene_id**-d, mille järgi antud sekvents oli genereeritud. **Gene_id**-d saadakse päringuga tabelist **Gene**.

Andmete päring

Olgu soov pärida kõiki geene, millel on seos motiiviga, kus esineb alamjärjestus

```
TCCGCTGAACCGTT
```

Esmalt pärime andmebaasis kõik sellised oligod, mis sisaldavad antud alamjärjestust. Ning seejärel pärime antud seondumissaitidega seotud geenid. Näidispäring:

```
Päring a:
SELECT  id, value, type, site_id
FROM  Signal
WHERE  value LIKE '%TCCGCTGAACCGTT%';
```

Tulemus a:

```
+-----+-----------------------------------------+-------+---------+
| id  | value                                   | type  | site_id |
+-----+-----------------------------------------+-------+---------+
| 158 | CGATGCGTCTTTTCCGCTGAACCGTT.    | oligo |     158 |
| 209 | gatGCGTCTTTTCCGCTGAACCGttc.    | oligo |     209 |
| 862 | GATGCGTCTTTTCCGCTGAACCGTTCCAGCAAAAAAGACTA | oligo |     862 |
+-----+-----------------------------------------+-------+---------+
```

Päring b:

```
SELECT  site_id, gene_id
FROM  Regulation
WHERE  site_id ='158'
OR site_id='209'
OR site_id='862';
```

Tulemus b:

```
+---------+---------+
| site_id | gene_id |
+---------+---------+
|     158 |       0 |
|     209 |       0 |
|     862 |      35 |
+---------+---------+
```

Päring c:

```
SELECT  id, name
FROM  Gene
WHERE  id='35';
```

Tulemus c:

```
+----+------+
| id | name |
+----+------+
| 35 | CUP1 |
+----+------+
```

3.4.3 Geen ja erinevad transkriptsiooni algussaidid

Kasutajal on identifitseeritud geen ja selle transkriptsiooni algussaidid (TSS). Eksisteerivad seosed:

- seondumissait on geeni ülesvoolu järjestuses

Andmete lisamine

Meetod ja autor kirjeldatakse tabelis **Source**. **Source** tabeli identifikaator lisatakse tabelitesse **Site**, **Signal, Gene** ning **Regulation**. Transkriptsiooni algussaidi (TSS) kirjeldused lisatakse tabelisse **Site**. Juhul kui TSS–i koordinaadid on samad, kuid nukleotiid on erinev, lisatakse iga nukleotiidi kohta kirje tabelisse **Signal**. **Value** atribuut saab väärtuseks antud nukleotiidi ning **type** on **TSS**. Juhul kui TSS–d on erinevate koordinaatidega, lisatakse iga TSS–i kohta üks kirje nii **Site** kui **Signal** tabelisse. Iga **Site** tabelisse kirje lisamisel luuakse **Regulation** tabelisse **site_id** ning **gene_id** väärtused. **Gene_id** saadakse päringuga **Gene** tabelist.

Andmete päring

Olgu soovitud geeni SPR3 transkriptsiooni algussaidid koos koordinaatidega. Näidispäring:

```
Päring a:
SELECT  id
FROM  Gene
WHERE  name='SPR3';

Tulemus a:
+-----+
| id  |
+-----+
| 158 |
+-----+

Päring b:
SELECT  Signal.value, Regulation.site_id
FROM  Regulation, Signal
WHERE  Regulation.gene_id='158'
AND Regulation.site_id=Signal.site_id
AND Signal.type='TSS';
```

Tulemus b:

value	coord_start	coord_end	site_id
G	-142	-142	1154
A	-147	-147	1155
G	-151	-151	1156
A	-163	-163	1157
A	-168	-168	1158
G	-173	-173	1159
C	-45	-45	1160
T	-58	-58	1161
C	-64	-64	1162
T	-65	-65	1163
T	-66	-66	1164
G	-67	-67	1165
T	-72	-72	1166
T	-73	-73	1167

Antud päringuga saame geenid, nende transkriptsioonialgussaidid (TSS) ja koordinaadid ORF–i suhtes.

Lisaks eeltoodud kolmele andmebaasis realiseeritud päringuvõimalusele kirjeldatakse veel kaht võimalikku kasutusjuhtu. Kuna alltoodud andmeid andmebaasis reaalselt ei eksisteeri, siis tuuakse vaid andmete lisamise kirjeldused.

3.4.4 Transkriptsioonifaktor ja ChIP on chip abil saadud geenid, kuhu antud transkriptsioonifaktor seondub

Kasutajal on kromatiini immuunosadestamise analüüsiga saadud andmed transkriptsioonifaktorite ning DNA komplekside moodustumise kohta. Eksisteerivad seosed:

- transkriptsioonifaktor seondub DNA–le

Andmete lisamine

ChIP on chip meetod ning autori andmed kirjeldatakse tabelis **Source**. Seejärel kirjeldatakse transkriptsioonifaktorit iseloomustavad tunnused tabelis

Gene ning **Factor**. Edasi lisatakse tabelisse **Regulation factor_id** ning **gene_id** –d, millele transkriptsioonifaktor seondub.

3.4.5 Klasterdamisel saadud *in silico* saidi kirjeldused

Kasutajal on geenieksрessiooni andmete analüüsil saadud sarnase ekspressioonimustriga geenide kogumid. Klasterdatud geenide ülesvoolu järjestustest on *in silico* analüüsidega leitud võimalikud transkriptsioonifaktorite seondumissaitide kirjeldused. Eksisteerivad seosed:

- seondumissait on geeni ülesvoolu järjestuses

Andmete lisamine

Esimeses etapis tuleb kirjeldata **Source** tabeli atribuutidega klasterdamismeetod, tõenäosuse lävi, andmete autor. Juhul kui lisatakse vaid üks regulaaravaldis, mis esitab saadud saite, siis piisab **Source** tabelis kirjeldatust. Kui lisatakse erineva skooriga ennustatud järjestusi, siis tuleb iga järjestuse kohta lisada uus **Site** tabeli kirje. Klasterdamisel saadud järjestus(ed) kirjeldatakse **Site** tabelis koordinaatidega, samuti märgitakse liigi nimi ning see, et andmed ei ole saadud eksperimentaalselt. Regulaaravaldise kujul olev järjestus lisatakse tabelisse **Signal**, **type** atribuut saab väärtuse **regular expression**. **Site** tabelisse kirjete lisamisega samaaegselt luuakse uued kirjed ka tabelisse **Regulation**, kus märgitakse ära milliste geenide eest on vastavad saitide saadud. **Gene_id**–d saadakse päringuga tabelist **Gene**.

3.5 Andmebaasi statistika

Seisuga 29.11.2003 on andmebaasis andmeid:

- kolmest eri allikast, seejuures kahest olemasolevast andmebaasist (Wingender *et al.* 2000; Zhu & Zhang 1999) ning lisaks (Kellis *et al.* 2003) artiklis avaldatud andmed.
- 783 geeni
- 219 faktorit
- 1291 saiti, nendest:
  - 1057 oligot
  - 0 konsensusjärjestust

- 195 transkriptsiooni algussaiti
- 39 maatriksit

Andmebaasi loomiseks ja andmete töötlemiseks on kirjutatud: üheksa andmetöötlus ning andmete sisestamise programmi, kogumahus 980 rida.

Kokkuvõte

Viimastel aastatel toimuv suuremahuline geeniregulatsiooni mehhanismide eksperimentaalne uurimine vajab andmebaaside toetust ning integreeritud *in silico* modelleerimise vahendeid. Mõistmaks geeniregulatsiooni võrgustikke tuleb teostada suuremahulisi andmeanalüüse ning selle toetuseks on vaja struktuurselt hästi modelleeritud ning bioloogilisi seoseid arvestavaid andmebaase.

Käesolevas töös anname ülevaate geeniregulatsiooni andmebaasist **BiGeR**, mis võimaldab integreerida erinevates juba eksisteerivates andmebaasides olevad andmed üheks tervikuks ja samas toetab uute andmete sisestamist ning analüüsi. Andmebaasi modelleerimiseks õppisime tundma geeniregulatsioonis osalevate bioloogiliste olemite omavahelisi seoseid ning neist lähtuvalt kujundasime andmebaasi struktuuri. Samuti lõime meetodid andmete töötlemiseks, millega erinevatest allikatest pärinevad andmed viiakse ühtsele kujule ning vastavusse meie poolt välja töötatud andmestruktuurile ning mille tagajärjel on erinevatest allikatest pärit andmed omavahel võrreldavad ning ühildatavad.

Antud töö teoreetilises osas andsime kirjanduse ülevaate geeniregulatsiooni mehhanismidest ning pikemalt käsitlesime transkriptsiooni ja selle kontrolli. Samuti kirjeldasime bioloogiliste andmebaaside peamisi omadusi. Teoreetilise osa meetodite pooles kirjeldasime *in vitro* ja *in silico* eksperimente transkriptsiooni seondumissaitide määramiseks.

Loodud andmebaas on kasutatav ka teiste organismide andmete mudeldamiseks. Projekti edasine eesmärk on arendada tööriistu nii arvutiprogrammide jaoks (programmeeritav kasutajaliides ehk API) kui ka tavaksutajate jaoks (veebipõhine). Edasine uurimustöö keskendub suuremahulisele geeniregulatsiooni andmete ja arvutuslike ennustuste võrdlemisele ning uute teadmiste genereerimisele.

Summary

Gene regulation at transcription level is the first and perhaps the most important step of the whole regulation machinery. Due advances in sequencing many complete DNA sequences can be studied for regulatory signals in the DNA. The aim of our work is to create a database for storing and analyzing data about gene regulation.

In this work we introduce BiGeR — a new database for storing gene regulation related information. The database gives us the possibility to analyze regulatory motifs in DNA and to compare different types of binding sites representations but also gives the chance to model gene regulatory networks and to study DNA motifs and their correlation.

We have populated the database with different data sources: gene regulation databases like Transfac (Wingender *et al.* 2000) and SCPD (Zhu & Zhang 1999), as well as *in silico* experiments and different articles which describe experimentally defined binding sites (Kellis *et al.* 2003).

Current work consists of two main parts, the theoretical, literature based overview, and the practical part about the design and usage of the database.

In the theoretical part of this work we describe control mechanisms of gene regulatory and mainly we introduce transcription regulation. In the methods chapter we show how gene regulation data can be analyzed – how it can obtained with *in silico* and *in vitro* experiments and how it is presented in different databases. We describe *in vitro* methods like *DNase I fingerprinting*, *mobility shift assay* and *chromatin immunoprecipitation*. We show also how regulatory regions can be defined with *in silico* methods like phylogenetic footprinting and gene expression data analysis. We studied different representations of transcription factor binding sites like oligos, matricies, consensus sequences and regular expressions. Also we show how regulation databases are made and give the basics for database modelling.

In the experimental part we describe the design of the database using the object model and table structure. Database functionality is described by several use cases and example queries.

Viited

Barash, Y.; Elidan, G.; Friedman, N.; and Kaplan, T. 2003. Modeling dependencies in protein-DNA binding sites. *RECOMB'03*.

Berg, O. G., and von Hippel, P. H. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology*.

Birney, E.; Clamp, M.; and Hubbard, T. 2002. Databases and tools for browsing genomes. *Annual Reviews Genomics Human Genetics* 3:293–310.

Blanchette, M., and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research* 12(5):739–748.

Brazma, A.; Jonassen, I.; Vilo, J.; and Ukkonen, E. 1998. Predicting gene regulatory elements in silico on a genomic scale. *Genome Research* 8:1202–1215.

Brown, T. 2001. *Gene Cloning and DNA analysis*. Blackwell Publishing, fourth edition.

Cao, D., and Parker, R. 2001. Computational modeling of eukaryotic mRNA turnover. *RNA* 7:1192–1212.

Charniak, E. 1991. Bayesian networks without tears. *AL. Magazine* 12(4).

Cliften, P.; Hillier, L.; Fulton, L.; Graves, T.; Miner, T.; Gish, W.; Waterston, R.; and Johnston, M. 2001. Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Research* 11:1175–1186.

Cockburn, A. Use case alternate intro.

Cornish-Bowden, A. 1985. IUPAC-IUB symbols for nucleotide nomenclature. *Nucleic Acids Research* 13:3021–3030.

Davis, C. A.; Grate, L.; Spingola, M.; and Ares, Jr., M. 2000. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Research* 28(8):1700–1706.

DeRisi, J.; Iyer, V.; and Brown, P. 1997. Exploring the methabolic and genetic control of gene expression on a genomic scale. *Science* 278:680–686.

Dwight, S.; Harris, M. A.; Dolinski, K.; Ball, C. A.; Binkley, G.; Christie, K. R.; Fisk, D. G.; Issel-Tarver, L.; Schroeder, M.; Sherlock, G.; Sethuraman, A.; Weng, S.; Botstein, D.; and Cherry, J. M. 2002. Saccharomyces

Genome Database (SGD) provides secondary gene annotation using the Gene ontology (GO). *Nucleic Acids Research* 30(1):69–72.

Ghosh, D. 1990. A relational database of transcription factors. *Nucleic Acids Research* 18(7):1749–1756.

Heumann, J.; Lapedes, A.; and Stormo, G. 1994. Neural networks for determing protein specificity and multiple alignment of binding sites. In *Conference on Intelligent Systems for Molecular Biology*, volume 2, 188–194.

Jenuwein, T., and Allis, C. D. 2001. Translating the histone code. *Science* 293:1074–1080.

Kang, S.-H. L.; Vieira, K.; and Bungert, J. 2002. Combining chromatin immunoprecipitation and DNA footprinting: a novel method to analyze protein-DNA interactions in vivo. *Nucleic Acids Research* 30(10):e44.

Kellis, M.; Patterson, N.; Endrizzi, M.; Birren, B.; and Lander, E.Š. 2003. Sequencing and comparison of yeast species to identify genes and regulatoru elements. *Nature* 423:241–254.

Kolchanov, N. A.; Ananko, E. A.; Podkolodnaya, O. A.; Ignatieva, E. V.; Stepanenko, I. L.; Kel-Margoulis, O. V.; Kel, A. E.; Merkulova, T. I.; Goryachkovskaya, T. N.; Busygina, T. V.; Kolpakov, F. A.; Podkolodny, N. L.; Naumochkin, A. N.; ; and Romashchenko, A. G. 1999. Transcription Regulatory Regions Database (TRRD): its status in 1999. *Nucleic Acids Research* 27(1):303–306.

Kolchanov, N. A.; Podkolodnaya, O. A.; Ananko, E. A.; Ignatieva, E. V.; Stepanenko, I. L.; Kel-Margoulis, O. V.; Kel, A. E.; Merkulova, T. I.; Goryachkovskaya, T. N.; Busygina, T. V.; Kolpakov, F. A.; Podkolodny, N. L.; Naumochkin, A. N.; Korostishevskaya, I. M.; Romashchenko, A. G.; and Overton, G. C. 2000. Transcription Regulatory Regions Database (TRRD): its status in 2000. *Nucleic Acids Research* 28(1):298–301.

Lee, T. I., and Young, R. A. 2000. Transcription of eukaryotic protein-coding genes. *Annual Review of Genetics* 34:77–137.

Liu, S.; Brutlag, D.; and Liu, J. 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin–immunoprecipitation microarray experiments. *Nature Biotechnology* 20(8):835–839.

Malik, S., and Roeder, R. G. 2000. Transcriptional regulation through mediator-like coactivators in yeast and metazoan cells. *Trends in Biochemical Sciences* 25:277–283.

Mulligan, M. E.; Hawley, D. K.; Entriken, R.; and McClure, W. R. 1984. *Escherichia coli* promoter sequences predict in vitro rna polymerase selectivity. *Nucleic Acids Research* 12:789–800.

Perier, R. C.; Praz, V.; Junier, T.; Bonnard, C.; and Bucher, P. 2000. The Eukaryotic Promoter Database(EPD). *Nucleic Acids Research* 28(1):302–303.

Praz, V.; Perier, R. C.; Bonnard, C.; and Bucher, P. 2002. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Research* 30(1):322–324.

Ptashne, M., and Gann, A. 1997. Transcriptional activation by recruitment. *Nature* 386(6625):569–577.

Qiu, P. 2003. Computational approaches for deciphering the transcriptional regulatory network by promoter analysis. *Biosilico* 1(4):125–133.

Ren, B.; Robert, F.; Wyrick, J. J.; Aparicio, O.; Jennings, E. G.; Simon, I.; Zeitlinger, J.; Schreiber, J.; Hannet, N.; Kanin, E.; Volkert, T. L.; Wilson, C. J.; Bell, S. P.; and Young, R. A. 2000. Genome-wide location and function of DNA binding proteins. *Science* 290:2306–2309.

Schneider, T. D.; Stormo, G. D.; Gold, L.; and Ehrenfeucht, A. 1986. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology* 188(3):415–431.

Staden, R. 1989. Methods for calculating the probabilities of finding patterns in sequences. *Computational Applications for Bioscience* 5:89–96.

Stormo, G., and Fields, D. 1998. Specificity, free energy and information content in protein-DNA interactions. *Trends in Biochemical Sciences.*

Stormo, G.; Schneider, T.; Gold, L.; and Ehrenfeucht, A. 1982. Use of the 'perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Research* 10:2997–3012.

Stormo, G. D. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16(1):16–23.

Tompa, M. 2001. Identifying functional elements by comparative DNA sequence analysis. *Genome Research* 11(7):1143–1144.

van Helden, J.; André, B.; and Collado-Vides, J. 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology* 281:827–842.

Vilo, J.; Brazma, A.; Jonassen, I.; Robinson, A.; and Ukkonen, E. 2000. Mining for putative regulatory elements in the yeast genome using gene expression data. In *Proceedings of the Eighth International Conference on Intelligent Syststems for Molecular Biology*, 384–394. AAAI Press, San Diego, CA.

Vilo, J. 2002. *Pattern Discovery from Biosequences.* Ph.D. Dissertation, University of Helsinki.

Weinmann, A., and Farnham, P. 2002. Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods* 26(1):37–47.

Wingender, E.; Chen, X.; Hehl, R.; Karas, H.; Liebich, I.; Matys, V.; Meinhardt, T.; Prüß, M.; Reuter, I.; and Schacherer, F. 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Research* 28(1):316–319.

www.whatislife.com. DNA binding proteins: nucleosome and transcription factors.

Zhang, M. Q. 2002. Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics* 3(9):698–709.

Zhu, J., and Zhang, M. Q. 1999. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* 15(7/8):607–611.