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

Bakalaureusetöö

Juhendaja: Jaak Vilo, PhD

Tartu 2004

Sisukord

Lühendid

A	Adenine	Adeniin
API	Application Programming Interface	Rakendusliides
Arg	Arginine	Arginiin
C	Cytosine	Tsütosiin
cDNA	complementary DNA	Komplementaarne DNA
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
HTML	HyperText Markup Language	Hüperteksti märgistuskeel
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
Perl	Practical Extraction and Reporting Language	Praktiline väljavõtte- ja aruandekeel
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 DNA-st bioloogiliselt olulisi signaale, toodetakse *in silico* ja *in vitro* uusi eksperimentaalseid ja ennustuslikke andmeid ning luuakse andmebaase saadud info esitamiseks. Selliste andmebaasside loomine on andnud võimaluse geeniregulatsiooni mehhanismide modelleerimiseks ja mõistmiseks, mis on tänapäeva molekulaarbioloogia suurimaid väljakutseid. Kuigi on loodud mitmeid erinevaid andmebaase geeniregulatsiooni andmete haldamiseks ja esitamiseks, puudus senini võimalus kompleksek andmete päringuks.

Käesolev töö annab esmalt kirjanduse põhise ülevaate geeniregulatsiooni mehhanismidest ja bioloogilistest andmebaasidest. Teoreetilises osas esitatakse bioloogiliste signaaljärjestuste erinevaid leidmis-, esitus- ja analüüsimeetodeid ning iseloomustatakse bioloogiliste andmebaaside modelleerimist.

Töö teises pooles antakse ülevaade valminud geeniregulatsiooni andmebaasist **BiGeR**. Täpsemalt käsitletakse *Saccharomyces 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 kasutuslugude ning näidispäringute kirjeldamine. Samuti tutvustatakse andmebaasi veebiversiooni prototüübi praeguseid võimalusi.

Peatükk 1

Kirjanduse ülevaade

Käesolevas peatükis käsitletakse esmalt geeniregulatsiooni erinevaid etappe ning detailsemalt peatutakse transkriptsioonil. Samuti antakse ülevaade transkriptsioonifaktoritest ja nende seondumissaitide leidmisest nii eksperimentaalselt kui arvutuslikult. Olulisel kohal on ka seondumissaitide esitusviiside kirjeldamine. Peatüki teises pooles tuuakse ülevaade andmebaaside modelleerimisest.

1.1 Geeniregulatsioon eukarüootsetes organismides

Geeniregulatsiooni teevad keerukaks organismi erinevad arenguetapid, pidev reageering keskkonna mõjudele, samuti erinevad rakutsükli etapid ning rakkude spetsialiseerumine. Kõigi eeltoodud faktorite mõju nõuab tugevat kontrollsüsteemi, mis võimaldaks määrata vajalike geenide ekspresseerimise. Organismide geeniregulatsioon on kompleksne mehhanism, mis koosneb mitmest erinevast etapist: transkriptsioon, transkriptsioonijärgne RNA protsessimine, RNA transport, mRNA stabiilsuse ehk eluea kontroll, translatsioon ning translatsioonijärgne protsessimine (Maimets 1999). Geeniregulatsiooni erinevate mehhanismide paremaks mõistmiseks on illustreeriv joonis 1.1.

1.1.1 Transkriptsioon eukarüootsetes organismides

Transkriptsioon on RNA süntees DNA kodeerivalt piirkonnalt, mida viib läbi RNA polümeraas. Antud töös käsitletakse mRNA sünteesi DNA–sõltuva RNA polümeraas II (pol II) poolt. Transkriptsioon on geeniregulatsiooni esmane etapp ning selle kontroll mängib geeni aktiivsuse regulatsioonil põhilist rolli. Transkriptsiooni initsiatsioon toimub geeni promootorregioonilt.

Promootorregioon sisaldab eukarüootidel TATA-boxi, GC elementi, mitmeid transkriptsioonifaktorite seondumissaite ja teisi regulaatorjärjestusi (Maimets 1999). Promootorregiooni pikkuseks loetakse kokkuleppeliselt kuni 10 000 aluspaari geeni algusest ülesvoolu (*upstream*) kõrgemate eukarüootide puhul või 600 aluspaari *S. cerevisiae* puhul. Promootoralale seonduvad TATA-boxile seonduv valk TBP (*TATA-binding protein*) ning sellega seotult 14-st valgust koosnev kompleks, mis seondub omavahel ja TBP–ga aga mitte DNA–ga. Samuti seostuvad promootoralale transkriptsioonifaktor IIB (TFIIB), mis seondub nii DNA kui pol II-ga (Buratowski 1996). Eelpoolmainitud valgud moodustavad basaalsed ehk transkriptsiooni initsiatsiooniks esmavajalike valkude hulga. Lisaks basaalsetele transkriptsioonifaktoritele on vaja ka spetsiifilisi transkriptsioonifaktoreid (Kimball 2001).

Spetsiifiliste transkriptsioonifaktorite seostumine DNA-ga on sõltuvuses rakutsükli etappidest, keskkonna mõjudest ning teistest ülalpool loetletud geeniregulatsiooni mõjutavatest faktoritest. Kui RNA sünteesi initsiatsiooniks vajalikud transkriptsioonifaktorid on promootoralale seondunud, toimub transkriptsiooni initsiatsioon 12–st valgust koosneva kompleksi ehk pol II seondumisega transkriptsiooni algussaidile (TSS).

Lisaks promootoraladele mõjutavad transkriptsiooni kontrolli ka spetsiaalsed regulaatoralad: võimendajad ehk *enhanser*–järjestused (*enhancer*), vaigistajad (*silencer*) ning piirielemendid ehk insulaator–järjestused (*insulator*) (Blackwood & Kadonaga 1998). Transkriptsioonifaktorid, seostudes nimetatud regulatoorsetele aladele, stimuleerivad või pärsivad transkriptsiooni. Võimendajate ülesandeks on transkriptsiooni efektiivsuse suurendamine, vaigistajate ülesanne on efektiivsuse vähendamine. Piirielemendid on DNA järjestused, mille ülesanne on reguleerida võimendajate ja vaigistajate ulatust, hoides ära võimendajate või vaigistajate interaktsioone promootoraladega. Võimendajaid ja vaigistajaid eristavad promootoritest järgnevad tunnused:

- võivad toimida tuhandete aluspaaride kauguselt
- mõju ei sõltu orientatsioonist
- mõju ei sõltu asukohast (võivad paikneda geenist eespool, tagapool või intronites)

Pikem ülevaade võimendajatest, vaigistajatest ning piirielementidest on toodud (Blackwood & Kadonaga 1998) artiklis.

1.1.2 Post-transkriptsiooniline mRNA protsessimine

Kui transkribeeritud RNA on juba pikem kui 30 aluspaari, lisatakse selle 5' otsa cap-struktuur. Cap-i ülesandeks on RNA kaitsmine degradatsioo-

ni eest ja transleerimise tagamine (*translatability*). Samuti vastutab 5' cap pre-mRNA transpordi eest tuumast tsütoplasmasse. mRNA-l võimendab 5' cap-i olemasolu transleerimist kuni 300-kordselt (Mallery 2004). Transkriptsiooni lõpetamiseks on vajalik polüadenülatsioon ehk polü-A saba lisamine pre-mRNA molekulile. Polü-A saba ülesandeks on samuti mRNA kaitsmine degradatsiooni eest ja samuti toimib see translatsioonisignaalina.

1.1.2.1 mRNA splaissimine

Valdav osa kõrgemate eukarüootsete organismide avatud lugemisraame (ORF-e) sisaldab introneid[1] . Transleeritava mRNA saamine toimub splaissosoomides, mille käigus lõigatakse intronid välja ja eksonid[2] ühendatakse.

1.1.2.2 Alternatiivne splaissimine

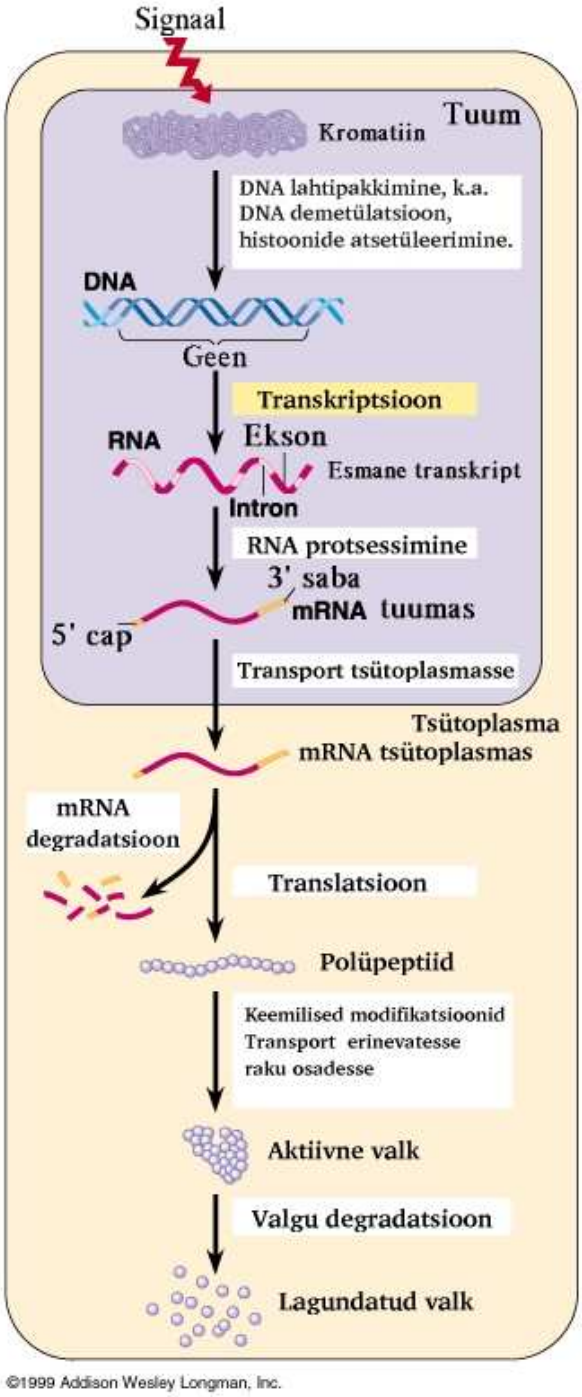
Mitmete intronite esinemine geenis võimaldab alternatiivset splaissimist. Alternatiivne splaissimine on protsess, mille käigus tekib ühest geenist mitu erinevat mRNA molekuli. Splaissimise käigus lõigatakse pre-mRNA-st lisaks intronitele välja ka eksoneid. Nii moodustuvad ühest geenist mitmed mRNA molekulid, 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).

1.1.2.3 mRNA stabiilsuse kontroll tsütoplasmas

Erinevalt eelkirjeldatud mehhanismidest, mis toimuvad tuumas, leiab valgusüntees ehk translatsioon aset tsütoplasmas. Transleeritava valgu hulka rakus on võimalik kontrollida mRNA hulga vähendamise või suurendamisega tsütoplasmas. Translatsioon mRNA molekulilt saab toimuda kuni mRNA molekuli degradeerumiseni. mRNA eluiga mõjutavad 3' polü–A saba olemasolu ja pikkus ning 3' mittetransleeritavad regioonid (Cao & Parker 2001).

[1]intron - geeni mittekodeeriv DNA järjestus, mis transkribeeritakse RNA molekuliks kuid lõigatakse sealt splaissimise käigus välja

[2]ekson - geeni kodeeriv DNA järjestus, millelt kodeeritakse mRNA



Joonis 1.1: Geeniregulatsiooni erinevad etapid
(Mallery 2004)

1.1.3 Transkriptsioonifaktorid

Geenide transkribeerimine on peamiselt seotud DNA-ga interakteeruvate valkude ehk transkriptsioonifaktorite äratundmise ning seostumisega. Transkriptsioonifaktorid seostuvad spetsiifilistele lühikestele DNA järjestusmotiividele ehk seondumissaitidele geenide *cis*-regulatoorses piirkonnas ja mõjutavad seeläbi otseselt või kaudselt transkriptsiooni toimumist. Transkriptsioonifaktorite seostumist spetsiifilisele DNA-le on võimalik täpselt määrata mitmete bioloogiliste ja keemiliste analüüsidega ning geeniregulatsiooni uurimisel enim mainituid käsitletakse peatükis 1.2.

1.1.4 Transkriptsioonifaktorite seondumissaidid

Genoomide sekveneerimine ning geeniregulatsiooni mehhanismide olulisus on motiveerinud teadlasi uurima transkriptsioonifaktorite seondumissaite. Seostumissaitide uurimine hõlmab kaht lähenemisviisi: bioloogiline *in vitro* ja *in vivo* meetodite näol ning arvustuslikud *in silico* analüüsid.

In vitro on võimalik DNA seostumisspetsiifikat kindlaks teha näiteks DNaas I jalajäljega(*DNase I footprinting*) ja elektromobiilse nihke analüüsiga(*shift assay*) (Qiu 2003). Lisaks mainitutele on veel mitmeid meetodeid (Nucleic Acid Research 2004), kuid neid käesolevas töös pikemalt ei käsitleta.

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

Viimased katsed on näidanud histoonivalkude modifitseerimise mõjusid transkriptsioonile. 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 olles hüperatsetüleeritud ja vastupidiselt võib olla transkriptsioonifaktori poolt mitte ära tuntav, kui genoomijärjestus on hüpoatsetüleeritud (Weinmann & Farnham 2002).

Mõlemad lähenemisviisid eeldavad transkriptsioonifaktorite seondumissaitides esineva motiivi kirjeldamist ja iseloomustamist. Transkriptsioonifaktorite seondumissaidid on 5-30 aluspaari pikad, seejuures enamik jääb va-

hemikku 5-16 aluspaari (Zhu & Zhang 1999; Vilo 2002; Qiu 2003). Seondumissaidi pikkuse määramisel tuleb arvestada, et sõltuvalt kasutatud eksperimentaalsest meetodist võib leitud järjestuse pikkus erineda (Zhu & Zhang 1999). Madalamate eukarüootide nagu *S. cerevisiae* organismis asuvad enamus regulatoorseid elemente 200–500 aluspaari ORFi 5' otsast ülesvoolu (Qiu 2003), kõrgemate eukarüootide puhul loetakse kokkuleppeliselt regulatoorseteks järjestusi, mis asuvad kuni 10 000 aluspaari geeni algusest ülesvoolu.

Levinumateks seondumissaitide esitusviisideks on oligonukleotiidid[3] ehk oligod ehk stringid, regulaaravaldised, maatriksid ning konsensusjärjestused. Nendest motiividest tuleb pikemalt juttu peatükis 1.4.

1.2 *In vitro* transkriptsioonifaktorite seostumissaitide määramine

Rakkudes toimuv geeniekspressiooni pidev ümberprogrammeerimine, mis on seotud rakutsükli ning keskkonnamuutustega, on põhjendatav DNA spetsiifiliste regulaatorite muutunud seostumisega. Erinevad DNA-ga seostuvad valgud interakteeruvad tsentromeeridega, telomeeridega ja teiste regulatoorsete piirkondadega. Seeläbi kontrollivad regulaatorvalgud kromosoomi replikatsiooni, kondensatsiooni, sidusust ja teisi genoomi säilitamise aspekte.

Ekspressiooniandmete analüüs DNA mikrokiibiga võimaldab uurijatel identifitseerida mRNA hulga suhtelisi muutuseid 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 (DeRisi, Iyer, & Brown 1997).

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

1.2.1 DNA–valk kompleksi liikuvuse muutus geelil

DNA–valk kompleksi liikuvuse (*retardation*) meetodi puhul valk-DNA interaktsiooni moodustumisel kasvab tekkiva kompleksi molekulaarmass ning

[3]Käesolevas töös kasutatakse oligonukleotiidide mõistet bioinformaatika terminoloogiale vastavalt tähendusega: mõne kuni mõnekümne aluspaari pikkune nukleotiidide järjestus

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 (Lane, Prentki, & Chandler 1992). Selliseks meetodiks sobib DNaas I jalajälg.

1.2.2 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.

DNaas I jalajälje meetodiga 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.

1.2.3 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 uuritav valk ei seostu DNA-ga kui guaniin on kontrollregioonis muudetud, kuna nukleotiidi metüleerimine segab spetsiifilist keemilist 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 guaniinidele, 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 uuritava geeni ekspressiooni muutusele. Kasutatakse reportergeene, mille ülesvoolu järjestusse kloneeritakse uuritava geeni promooterala. Kloonitult peaks reportergeeni ekspressiooni profiil täpselt jäljendama originaalgeeni, kui reporter geen on täpselt sama kontrolljärjestuse mõju all kui originaalgeen. Reportergeeni valimisel tuleb jälgida, et geeniga kaasnev fenotüüp ei tohi olla juba avaldunud peremees organismis, et fenotüüpi oleks kerge detekteerida ja kui võimalik, siis oleks fenotüüpi võimalik kvantitatiivselt mõõta. Kloneeritud reportergeeni ülesvoolu järjestustest lõigatakse võimalikke regulaatoralasid. Seejärel viiakse muudetud konstrukt peremeesorganismi ning jälgitakse 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 eemaldumisele.

Eelnevalt kirjeldatud meetodid sobivad juba piiritletud regulaatorregioonide täpsemaks uurimiseks ning seondumissaitide täpseks määratlemiseks. 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 teostatakse suuremahulisi valk-DNA interaktsioonide uuringuid on kromatiini immunosadestamine. Järgnevalt käsitletakse seda pikemalt.

1.2.4 Kromatiini immunosadestamine kiibil

Kromatiini immunosadestamine kiibil (*Chromatin immunoprecipitation* (ChIP on Chip)) võimaldab jälgida valk–DNA interaktsioone üle terve genoomi ning seeläbi võimaldab leida transkriptsioonifaktorite seostumisi *in vivo* ja seeläbi analüüsida regulatoorseid võrke (Qiu 2003; Ren *et al.* 2000). Genoomis esinevate seostumisandmete ja ekspressiooniandmete kombinatsioon võimaldab identifitseerida üldist geenide hulka, mille ekspressioon rakus on otseselt kontrollitud transkriptsiooni aktivaatorite poolt (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. Kromatiini immunosadestamise meetodi peamine puudus seisneb selles, et antikehadega rikastatud DNA seostumine regulaatorvalkudega ei viita alati sellele, et valk seostub sadestatud järjestusega. Meetod võib viidata hoopis valk-valk interaktsioonidele (Kang, Vieira, & Bungert 2002). Kromatiini immunosadestamine kiibil suudab määrata seostumise ühe- kuni kahetuhande aluspaari täpsusega (Liu, Brutlag, & Liu 2002).

Rakud fikseeritakse formaldehüüdiga, kogutakse ja töödeldakse ultraheliga. DNA fragmente, mis on ristseotud(*cross-linked*) meid huvitava valguga märgitakse 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 viiakse läbi ligeerimisvahendatud polümeraasi ahelreaktsioon. Nii immunosadestamisel rikastatud kui ka rikastamata proovid hübridiseeritakse DNA mikrokiibile, mis sisaldab kõiki uuritava organismi 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).

Kromatiini immunosadestamise 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).

1.3 *In silico* analüüs

Bioloogilised meetodid transkriptsioonisaitide määratlemiseks on ühelt poolt ebatäpsed (kromatiini immunosadestamine kiibil) ning teisalt liiga töömahukad ning suure ajakuluga (DNaas I jalajälg). Meetodite, mis suudavad detekteerida valk-DNA seondumisi, tulemused on aga headeks lähteandmeteks suure analüüsivõimega *in silico* meetoditele. Arvutuslike meetodite peamine eesmärk on analüüsida teadaolevat bioloogilist infot ning ennustada uusi võimalikke regulaatorpiirkondi. Kui *in silico* meetoditega on määratletud kandidaatregioonid on võimalik neid bioloogiliste meetoditega kontrollida. Seega on bioloogiliste eksperimentide tulemused heaks lähtebaasiks arvutuslikele analüüsidele ning *in silico* meetodid omakorda võimaldavad bioloogilisi eksperimente paremini planeerida.

In silico transkriptsioonifaktorite seondumissaite ennustades on kaks peamist ülesande püstitust.

Esimene ülesanne on ennustada tõenäolisi seostumissaite juba tuntud transkriptsioonifaktorile üle genoomi. Sellisel juhul kasutatakse teadaolevaid seondumissaitide näiteid ning nende põhjal genereeritud mudeliga otsitakse genoomist uusi mudelile vastavaid võimalikke saitide esinemisi.

Teine ülesanne on teadmata transkriptsioonifaktori seondumissaiti leida tõenäoliselt sama faktori poolt reguleeritud geenid. Nende ülesvoolu järjestustest tuleb otsida transkriptsioonifaktori seondumismotiiv ning seejärel leida saadud motiivi esinemised ka mujal genoomis.

Tavaliselt kasutatavad motiivide esitusviisid on positsiooni spetsiifiline skoorimaatriks (PSSM) ning Rahvusvaheline Puhta ja Rakenduskeemia Liidu poolt välja töötatud IUPAC koodi (Cornish-Bowden 1985) kasutav PROSITE tüüpi mustrid. Need mustrid on oma olemuselt regulaaravaldised, mis võimaldavad kirjeldada ühetäheliste kombinatsioonidega lisaks neljale nukleiinhappele ka näiteks üldisemalt puriine, pürimidiine - vastavalt R ja Y. 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 võimalikke transkriptsioonifaktorite seondumissaite on saadud *in silico* ennustamisel. Tänapäeval teostatakse peamiselt *in silico* ennustusi geenikspressiooniandmete analüüsil ja fülogeneetilisel jalajäljel põhinevate meetoditega.

1.3.1 Geenieksрressiooni andmete analüüs

Geenieksрressiooni andmete analüüs põhineb DNA kiibi tehnoloogial. Mikrokiibi tehnoloogia mRNA populatsiooni suhtelise hulga mõõtmiseks rakkudes võimaldab meil jälgida tuhandete geenide ekspressiooni tasemeid üheaegselt. Mõõtes mitmetel erinevatel tingimustel või ajamomentidel ekspressiooni lävesid, on võimalik geenieksрressiooni kaardi koostamine (Brazma *et al.* 1998).

Geenieksрressiooni katse eksperimentaalses osas *in vitro* seotakse kiibile ülesamplifitseeritud 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).

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

1.3.2 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).

1.3.2.1 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).

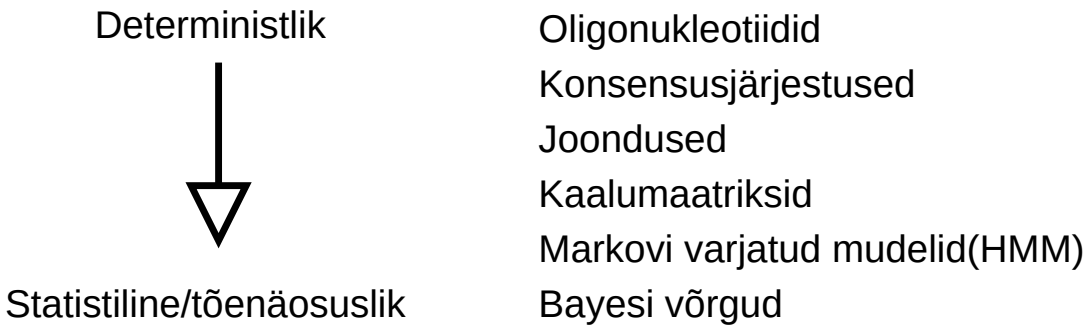
Genoomid tuleb valida põhimõtte järgi, et nad ei oleks liiga lähedased ega liiga kauged evolutsioonilises puus. Liiga lähedaste genoomide puhul suudetakse küll järjestused hästi joondada, kuid funktsionaalsed elemendid ei ole märgatavalt paremini konserveerunud ning seega ei saa neid eristada mittefunktsionaalsetest 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).

Fülogeneetilise jalajälje meetodist annab pikema ülevaate uurimistöö (Peterson 2004).

1.4 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 oligonukleotiidide, 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 1.2: Transkriptsioonisaitide esitusviiside jaotus deterministlike ja statistiliste vahele. Nool näitab suunda lihtsamatelt ja väiksema väljendusrikkusega esitusviisidelt keerulisemate ja suurema väljendusvõimsusega esitusviiside poole.

1.4.1 Oligonukleotiidid

Oligonukleotiidi tüüpi avaldised on informaatika mõistes täpsed alamstringid ehk alamsõned, mis pärinevad enamasti eksperimentaalselt tõestatud andmetest. Samuti on oligonukleotiidi kujul esitatud enamik andmebaasides olevaid seondumisaite.

Reeglina oligonukleotiididega otsitakse täpseid esinemisi ehk otsimisel leiab ainult 100% sama oligonukleotiidi. Sellise otsimise miinuseks on, et ei leita otsitavale sarnaseid järjestusi. Ligikaudsete stringide otsimise lahenduseks on teisenduskauguse kasutamine. Teisenduskaugus väljendab minimaalset arvu teisendusi, mida on vaja, et saada oligonukleotiidist A oligonukleotiid B. Vähimat arvu insertsioone, deletsioone või asendusi, mida läheb vaja oligonukleotiidi A teisendamiseks oligonukleotiidiks B nimetatakse Levenshteini kauguseks. Bioloogias kasutatakse muutmiskaugust sageli kahe organismi evolutsioonilise kauguse hindamiseks.

1.4.1.1 Näidisoligonukleotiidid

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

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

Levenshteini kaugus näites 1.4.1.1 toodud viienda (TCACCT) ja kuuenda (TCCGGT) oligonukleotiidi vahel on 3 ühikut.

1.4.2 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 1.1 nukleotiidide mitmesuse esitamiseks (Cornish-Bowden 1985):

Tabel 1.1: Laiendatud DNA / RNA tähestik

Sümbol	**Tähendus**	**Nukleiinhape**
A	A	Adeniin
C	C	Tsütosiin
G	G	Guaniin
T	T	Tümiin
U	U	Uratsiil
M	A või C	
R	A või G	Puriinid
W	A või T	
S	C või G	
Y	C või T	Pürimidiinid
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	

1.4.2.1 Näidis konsensusjärjestus

Peatükis 1.4.1.1 kirjeldatud näidisoligonukleotiidide põhjal leitud konsensusjärjestus:

```
TCCGCT   (kui igas positsioonis kirjeldatakse
          ainult enim esinenud nukleotiid)

WMMSSW   (kui igas positsioonis kirjeldatakse
          IUPAC'i tähestikku kasutades ära kõik esinenud
          nukleotiidid)

TMMGST   (kui igas positsioonis kirjeldatakse
          enim esinenud nukleotiid (kui esinemissagedus
          võrreldes järgmise nukleotiidi sagedusega on
          suurem kui 2) või IUPAC'i tähestikku kasutades
          võimalikud nukleotiidid (kui enim esinenud
          nukleotiid esineb 2 või vähem korda enam kui
          järgmine nukleotiid))
```

1.4.3 PROSITE tüüpi regulaaravaldised

Regulaaravaldised, mida kasutatakse bioloogiliste saitide kirjeldamiseks, on vaid osa üldistest regulaaravaldistest. Enamasti bioloogias kasutatakse PROSITE tüüpi mustreid. PROSITE esitusviis kasutab IUPAC-i ühetähelisi koode. Kuigi PROSITE mustrid on mõeldud peamiselt valkude kirjeldamiseks, siis sarnast esitusviisi kasutatakse ka nukleotiidide puhul. PROSITE tüüpi regulaaravaldised sisaldavad järgmisi omadusi:

IUPAC koodi tähised aminohapete ning nukleotiidide kirjeldamiseks

[] kirjeldamaks lubatud nukleotiide, näiteks [AC] puhul on lubatud adeniin ja tsütosiin.

{} kirjeldamaks mitte lubatud nukleotiide, näiteks {CG} puhul ei ole lubatud tsütosiin ja guaniin.

() abil kirjeldatakse lubatud nukleotiidide kordusi, näiteks T(3) puhul on kolm järjestikkust tümiini.

Näites 1.4.1.1 kirjeldatud oligonukleotiidide põhjal genereeritud PROSITE tüüpi 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.

Regulaaravaldisi kasutatakse peamiselt UNIX operatsioonisüsteemis ning mitmetes programmeerimiskeeltes, näiteks Perlis, ligikaudsete stringide otsimiseks tekstist. Informaatikaliselt on võimalik kirjeldada näites 1.4.1.1 toodud oligote põhjal regulaaravaldis [TA][CA][^GT][GC].[TA], kus [^GT] tähistab kõiki nukleotiide peale G ja T ning punkt tähistab ükskõik millist nukleotiidi.

Samuti kasutatakse informaatikas regulaaravaldiste puhul sümboleid *, + ja ?. Sümboli * tähendus on null või rohkem tärnile eelnevat sümbolit, + tähendab vastavalt 1 või rohkem korda ning ? tähendus on null või üks korda eelnevat sümbolit.

Regulaaravaldisele A*T?G+ vastavad stringid, mille alguses võib olla piiramatu hulk (kaasa arvatud mitte ühtegi) A sümbolit, seejärel üks või mitte ühtegi T sümbolit ning lõpus vähemalt üks G.

1.4.4 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.

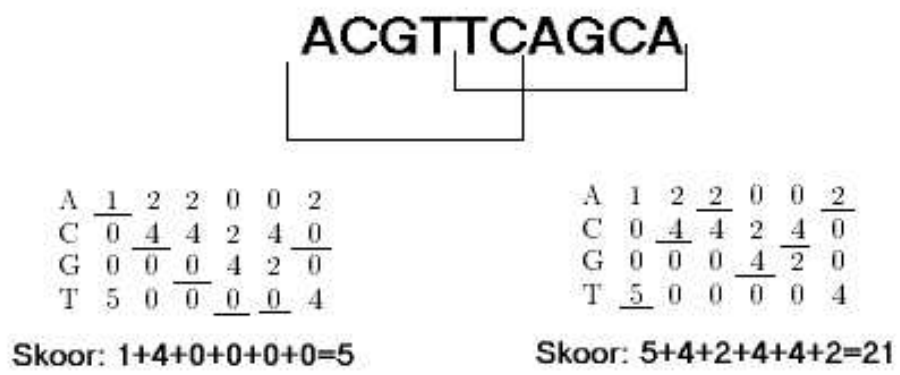
Kaalumaatriks (*Position weight matrix, PWM*) on alternatiiv konsensusjärjestusele. Esmalt kasutati kaalumaatrikseid RNA saitide iseloomustamiseks, mis funktsioneerisid *E.coli* translatsiooni initsiatsioonisaitidena (Stormo *et al.* 1982). Leiti, et lisaks Shine-Dalgarno järjestusele ning initsiatsioonikoodonile on ka ribosoomi seostumissaidid kõrgelt konserveerunud (Stormo 2000).

Sellest järeldub, et mitmed aluspaarid ribosoomi seostumisregioonis mRNA-l võivad interakteeruda ribosoomiga ja tõenäosus, et seostumine on piisav initsieerimaks 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).

Tabel 1.2: Näites 1.4.1.1 esitatud nukleotiidide sageduste järgi 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

Maatriksite sobitamisel stringidele võrreldakse iga positsiooni stringis vastava positsiooniga maatriksis ja leitakse kaalud. Joonisel 1.3 on kirjeldatud tabelis 1.2 esitatud positsioonimaatriksi sobitamine stringile ACGTTCAGCA. Sobitades etteantud maatriksi 1.2 stringi algusest kuue aluspaari ulatuses, saame skoorid vastavalt: esimeses positsioonis A = 1, teises positsioonis C = 4, ülejäänud positsioonides vastavalt G, T, T, C on kaal 0. Kokku annab maatriksi sobitamine stringile positsioonides üks kuni kuus kaaluks viie. Sobitades sama maatriksi stringile positsioonist viis kuni positsioonile kümme, saame skooriks 19. Parima sobivuse puhul oleks skoor 25.



Joonis 1.3: Maatriksi 1.2 sobitamine stringile ACGTTCAGCA. Esmalt sobitatakse maatriks stringile positsioonidele 1 kuni 6. Tulemuseks on sobivust väljendav skoor väärtusega 5. Sobitades maatriksit stringile positsioonidele 5 kuni 10 on skooriks 21.

1.4.4.1 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).

On näidatud, et järjestuse skoori ja promootori aktiivsuse vahel eksisteerib tugev korrelatsioon. 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 sõl-

tumatud. Limiteerivaks on kvantitatiivse andmestiku saamise töömahukus ning seepärast kasutatakse sellist lähenemist väga harva (Stormo 2000).

1.4.4.2 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:

$$I_i = 2 + \sum_{b=A}^{T} f_{b,i}\, log_2 f_{b,i} \tag{1.1}$$

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:

$$I_{seq(i)} = \sum_{b} f_{b,i}\, log_2 \frac{f_{b,i}}{p_b} \tag{1.2}$$

kus p_b on aluse b sagedus kogu genoomis. Valem 1.1 on valemi 1.2 erijuht, kus p_b on kõikide b jaoks 0.25. I_{seq} on tuntud kui suhteline entroopia ja Kullback-Liebler kaugus.

Tabel 1.3: Informatsiooni sisalduse maatriks näite 1.4.1.1 põhjal arvutatuna valemi 1.2 järgi

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

1.4.5 Markovi varjatud mudelid

Markovi varjatud mudelid (*Hidden Markov model, HMM*) kirjeldavad süsteemi, mis koosneb eraldiseisvatest olekutest ja olekute vahelistest seostest. Iga seost iseloomustab tõenäosus. Mudelid on "varjatud", kuna seisundeid ei saa otseselt jälgida. Markovi varjatud mudelid üldistavad eelnevalt kirjeldatud positsioonimaatrikseid, sest nad võtavad arvesse eelmiste veergude seisundeid. Bioinformaatikas on HMM oluline seetõttu, et võimaldab otsida või luua joonduse algoritmi kindla tõenäosuse baasil ja mudelit on lihtne treenida tuntud andmestikuga (Zhang 2002). Markovi varjatud mudelid arvestavad tõenäosuse arvutamisel ka eelnevas positsioonis oleva nukleotiidi väärtust ja seega on statistiliselt väljendusrikkamad kui positsioonimaatriksid.

1.4.6 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 eelnevatest 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).

1.5 Bioloogilist infot sisaldavad andmebaasid

Bioloogiliste eksperimentide tulemusena tekkivad andmehulgad vajavad säilitamismooduseid, mis võimaldaks kirjeldada saadud tulemused võimalikult täpselt ilma bioloogiliselt olulist infot kaotamata.

1.5.1 Bioloogiliste andmebaaside vajadused

Maksimaalse kasu saamiseks bioloogilistest eksperimentidest pärit andmete rohkusest, tuleb need andmehulgad 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, RNA-st, 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 silmas 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 (Elmasri & Navathe 2000):

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

1.5.2 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 dünaamilised, 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.

1.5.3 Geeniregulatsiooni andmebaasid

Geeniregulatsioon hõlmab kõiki 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 (Kolchanov *et al.* 1999; 2000).

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

1.6 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 (Elmasri & Navathe 2000).

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 (Elmasri & Navathe 2000). ER mudeli eesmärgiks on andmebaasi kontseptuaalne kirjeldamine. Relatsioonilise andmemudeli puhul on keskseks objektide väärtustest lähtumine. 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.

1.6.1 Relatsioonilised andmebaaside haldamise süsteemid

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.

1.6.2 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 (Elmasri & Navathe 2000). Samas tekitab relatsiooni-

de 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.

Tabel 1.4: Inimesed

Perekonnanimi	Eesnimi	Isikukood
Mets	Mari	47712030987
Meri	Meelis	36709181202

Tabel 1.5: Kontoomanikud

Isikukood	Kontonumber
47712030987	12345656
36709181202	23423423

Olgu meil näiteks andmebaas, mis koonseb kahest tabelist 1.4 ja 1.5. Esimene tabel vastab olemite klassile *Inimesed*. Iga rida tabelis 1.4 vastab ühele objektile ehk antud juhul inimesele. Iga objekt on kirjeldatud kolme väljaga. Objekti tunnusteks on antud juhul *Perekonnanimi*, *Eesnimi* ja *Isikukood*. Tabelis 1.5 on kirjeldatud olemite klass *Kontoomanikud*. Sellesse olemite klassi kuuluvaid objekte iseloomustavateks väljadeks on *Kontonumber* ja *Isikukood*.

1.6.3 Võtmed

Võti on atribuut ehk omadus või ka atribuutide kogum, mis üheselt määrab ära konkreetse olemi (Elmasri & Navathe 2000). Ühel olemil võib olla mitu võtit kuid enamasti määratakse üks võtmete seast primaarvõtmeks. Ülejäänud võtmed on kandidaatvõtmed. Supervõti on atribuutide hulk, mille pärisalamhulk[4] ei ole võti.

Kasutades näitena tabeleid 1.4, 1.5 võime olemite klassi *Inimesed* primaarvõtmena käsitleda välja *Isikukood* kuna see on unikaalne numbrikombinatsioon ehk ei leidu kahte inimest, kellel oleks sama isikukood. Tabeli 1.4 primaarvõti on välisvõtmeks tabelis 1.5 ehk välja *Isikukood* abil saame me siduda kaks tabelit *Inimesed* ja *Kontoomanikud*.

Olemi kogumid, millel puudub võti, nimetatakse nõrkadeks olemikogudeks. Tugev olemikogu omab primaarvõtit.

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

1.6.4 Olem-seos mudel

Olem-seos (ER) mudel on vajalik reaalse maailma kirjeldamiseks ning olemite ja seoste määratlemiseks, enne kui asutakse modelleerima andmebaasi (Elmasri & Navathe 2000). 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.

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

1.6.5 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 päringukeel(DQL) –kasutatakse andmete pärimiseks andmebaasist.

  SELECT lause ehk kirjete pärimine

Relatsioonilised andmebaasid kasutavad peamiselt SQL keelt ning kõik päringud andmebaasist toimuvad SELECT lausena.

1.6.6 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 (Elmasri & Navathe 2000):

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.

Atomaarsuse parimaks näiteks võib tuua pangaülekanded. Transaktsiooniks on sel juhul esmalt raha võtmine kontolt A ning seejärel kandmine kontole B. Vajalik on teostada mõlemad toimingud järjest ning ilma katkestuseta ehk kui raha on võetult kontolt A, siis peab see kantama kontole B. Transaktsiooni alustatakse esimese SQL lause täitmisel ning lõpetatakse:

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

1.6.7 Teoreetilise osa kokkuvõte

Geeniregulatsioon on keeruline mehhanism ning selle uurimiseks tuleb esmalt tunda transkriptsioonimehhanisme. Selleks, et teada millised interaktsioonid tekivad transkriptsioonifaktorite ja DNA vahel tuleb teostada mitmeid eksperimentaalseid katseid. Samas on katsete täpsus olenevalt eksperimendist väga erinev ning samuti on bioloogiliste eksperimentide ajakulu väga suur ning pole võimalik kõiki võimalikke transkriptsioonifaktoreid ja DNA vahelisi interaktsioone eksperimentaalselt uurida.

Siinkohal tuleb appi bioinformaatika, mis võimaldab teha suuremahulisi *in silico* eksperimente ennustamaks transkriptsioonifaktorite seostumissaite DNA-l ning võimalike geeniregulatsiooni võrgustikke.

Samas nõuavad nii bioloogilistes eksperimentidest kui arvutuslikest analüüsidest tulevad andmed säilitamist, kirjeldamist ning analüüsimist. Selleks on vaja andmeid koguda ning hoida spetsiaalselt selleks tarbeks modelleeritud andmebaasides. Andmebaaside modellerimisel tuleb silmas pidada kirjeldavate bioloogiliste andmete olemust ning omavahelisi seoseid. Samuti on oluline et andmebaas toetaks nii *in vitro* ja *in vivo* kui ka *in silico* eksperimentide tulemuste esitamist. Ühist ja võrreldavat esitamist vajavad ka erinevad transkriptsioonifaktorite seondumissaitide esitusviisid, olgu need siis oligod, maatriksid või regulaaravaldised.

Geeniregulatsioon on väga tähtis mehhanism organismides ning selle uurimine on tänapäeva bioloogias olulisel kohal ning nõuab spetsiaalselt geeniregulatsiooni vajadustele disainitud andmebaase. Järgnev peatükk annab ülevaate esimesest võimalikust lahendusest kirjeldamaks geeniregulatsiooni infot spetsiaalselt selleks otstarbeks disainitud andmebaasist **BiGeR**.

Peatükk 2

Geeniregulatsiooni andmebaas BiGeR

2.1 Ülesande püstitus

Geeniregulatsiooni mehhanismide mõistmiseks on vajalik omada transkriptsioonifaktorite ning DNA interaktsioonide kohta infot. Tänapäeval saadakse geeniregulatsiooni andmeid peamiselt kahel viisil: arvutuslikest ennustustest ehk *in silico* ning bioloogilistest *in vitro* ja *in vivo* eksperimentidest. Saadud andmed tuleb töödelda ning esitada parimat modelleerimist võimaldaval kujul. Geeniregulatsiooni kirjeldavate andmete üha suurenev hulk vajas spetsiaalset andmebaasi, mis oleks disainitud ühendamaks erinevatest allikatest pärit infot ning võimaldamaks geeniregulatsiooni võrgustike modelleerimist. Käesoleva töö eesmärgiks oli luua andmebaas, mis vastaks eelpool mainitud nõudmistele ning oleks võimalikuks alusepanijaks eksperimentaalsete ja ennustuslike andmete koosesitamisele ning võrdlemisele.

2.2 Tulemused

Olles esmalt tundma õppinud bioloogilisi andmeid kui olemeid ning andmete omavahelisi seoseid, tuli võimalikult hästi püüda edasi anda neid seoseid ning olemeid ka andmebaasis. Oluline oli 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, kus transkriptsioonifaktor seondub DNA–le ning seda seondumissaiti on võimalik esitada nukleotiidide järjestusena
- motiiv on kogum, mis esitab ühe transkriptsioonifaktori seondumissaitides 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 juba olemasolevate andmete hoidmiseks ning pärimiseks vaid peamiselt edaspidiste *in silico* eksperimentide toetuseks. Andmebaasi loomisel oli eesmärgiks olla informaatiliseks baasiks geeniregulatsiooni kirjeldamisele, uurimisele ja võimaldada geeniregulatsioonivõrgustike modelleerimist.

2.2.0.1 Töö peamised etapid

Käesoleva uurimistöö peamised etapid on olnud:

1. andmebaasi struktuuri väljatöötamine
2. andmete kogumine
3. andmete töötlemine ühtsele kujule
4. andmebaasi programmeerimine
5. andmete automaatne sisestamine andmebaasi
6. veebiliidese programmeerimine

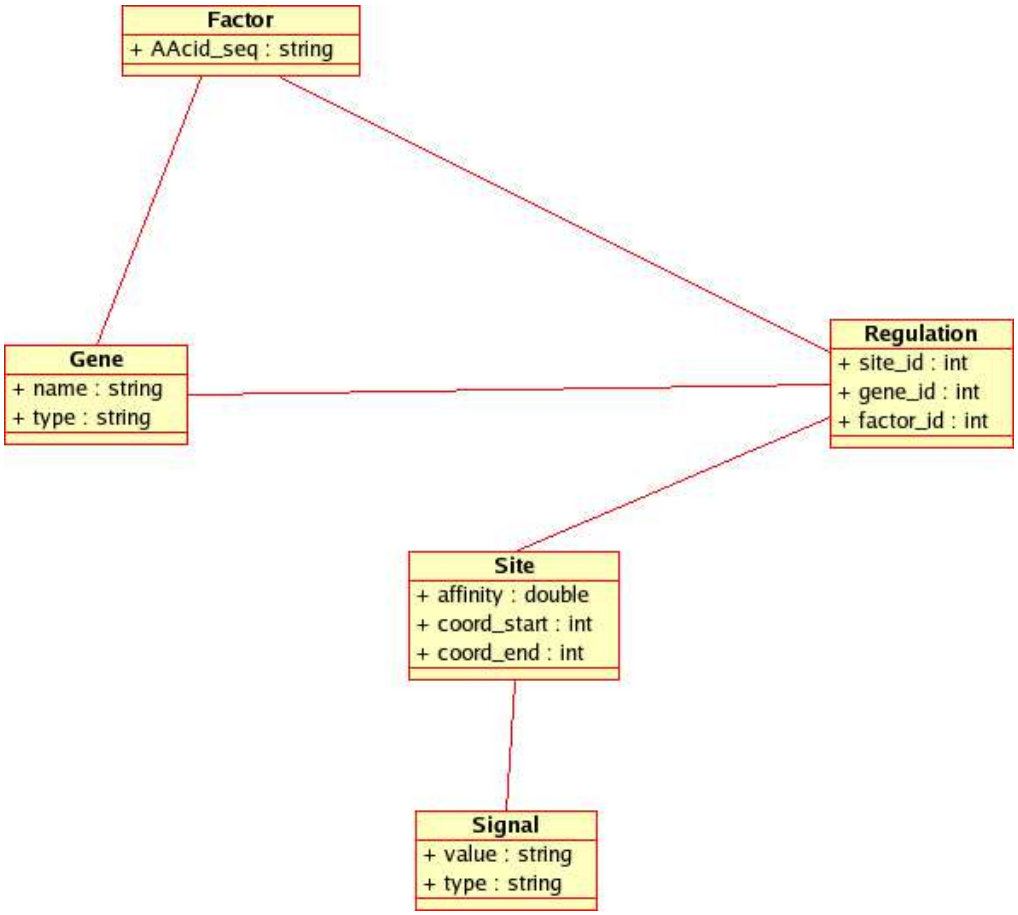
Etappe üks kuni kolm tuli korrata, et leida parim lahendus olemasolevate ja võimalike uute andmete hoidmiseks ja kirjeldamiseks. Andmebaasi struktuuri väljatöötamise peamiseks ja töömahukaimaks etapiks oli geeniregulatsiooni, peamiselt transkriptsiooni, mehhanismide kirjeldamine informaatiliste seostena.

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.

[1] Bioinformatics of Gene Regulation

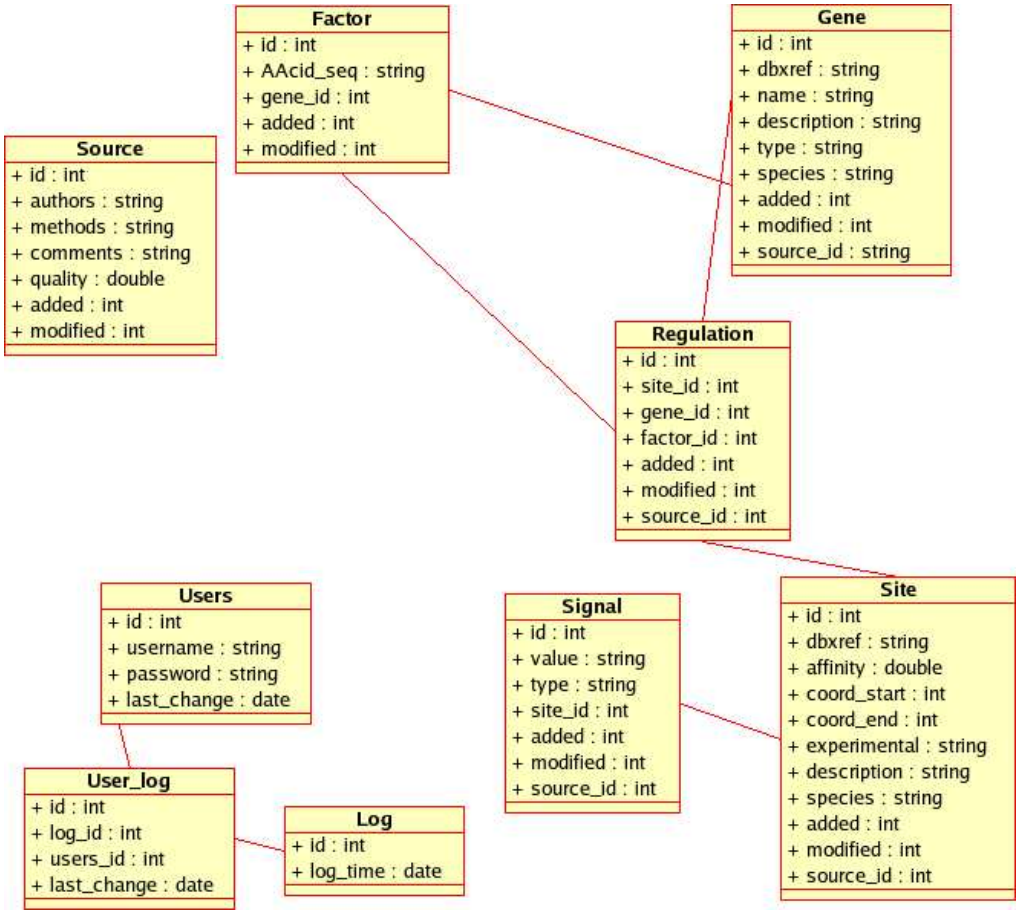
2.3 Andmebaasi skeem

Andmebaasi struktuuri on käesolevas peatükis kirjeldatud kolme joonise abil: joonis 2.1 annab ülevaate **BiGeR** andmebaasi peamistest olemitest ja nende kõige olulisematest atribuutidest, joonis 2.2 kirjeldab ülevaatlikult kõiki andmebaasi olemeid koos kõigi atribuutidega, lisaks on joonisel 2.3 välja toodud kasutajate haldamiseks vajalike tabelite skeem.



Joonis 2.1: **BiGeR** peamised tabelid olulisemate argumentidega. Tabel **Gene** kirjeldab geenide üldisi omadusi, **Factor** kirjeldab lisaks transkriptsioonifaktoritele olulisi omadusi. **Site** kirjeldab seondumissaitide omadusi ning tabelis **Signal** hoitakse seondumissaitide DNA järjestusi. Tabel **Regulation** kirjeldab geene ja neid mõjutavaid transkriptsioonifaktoreid ning samuti seondumissaite, mis asuvad geenide ees ja vastavad transkriptsioonifaktoritele.

Andmebaasi lihtsustatud mudeli joonisel 2.1 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 transkriptsioonifaktorid nende seondumissaitidega. Selline ühtse tabeli kujul geeniregulatsiooni modelleerimiseks vajaminevate andmete esitamine on autorile teadaolevalt senini olemasolevates andmebaasides puudunud.



Joonis 2.2: **BiGeR**-i täielik objekt mudel. Siin on ära toodud tabelite kõik atribuudid ning lisatabelid: **Source**, **User**, **Log** ja **User_log**.

Lisaks joonisel 2.1 toodud lihtsustatud mudeli viiele tabelile on olulisel kohal täieliku mudeli joonisel 2.2 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**.

2.4 Andmebaasi klasside detailsed 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.

2.4.1 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.

2.4.2 Tabel Gene

Gene kirjeldab geenide ehk valku kodeerivate DNA järjestuste lihtsamaid omadusi. Tabeli eesmärgiks ei ole koguda kõiki teadaolevaid andmeid iga geeni kohta vaid pigem salvestada viited erinevatele välistele andmebaasidele.

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**).

2.4.3 Tabel Factor

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

Atribuudid

AAcid_seq on aminohapete järjestus oligonukleotiidi kujul.

2.4.4 Tabel Site

Klass **Site** kirjeldab transkriptsioonifaktori seondumissaite DNA–l ning transkriptsiooni algussaite (TSS). Tabelis kirjeldatakse saidid suhteliste koordinaatidega geeni avatud lugemisraami algusest. Klass **Site** on otseselt seotud klassiga **Signal**, mis kirjeldab seondumissaitide DNA järjestusi.

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.

2.4.5 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**.

2.4.6 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. **Regulation** on peamiseks allikaks geeniregulatsiooni võrgustike modelleerimisel ning on keskseks tabeliks andmebaasis **BiGeR**.

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**

2.4.7 Tabel Source

Source on klass andmete päritolu kirjeldamiseks ning on aluseks andmete kvaliteedi hindamise väljatöötamisel. Tabel on loodud eesmärgiga võrrelda erinevatest allikatest pärinevate andmete kvaliteeti ning seeläbi töötada välja andmete usalduspiirid. Näiteks eksperimentaalsed andmed on suurema usaldusväärsusega kui arvutuslikud ennustused.

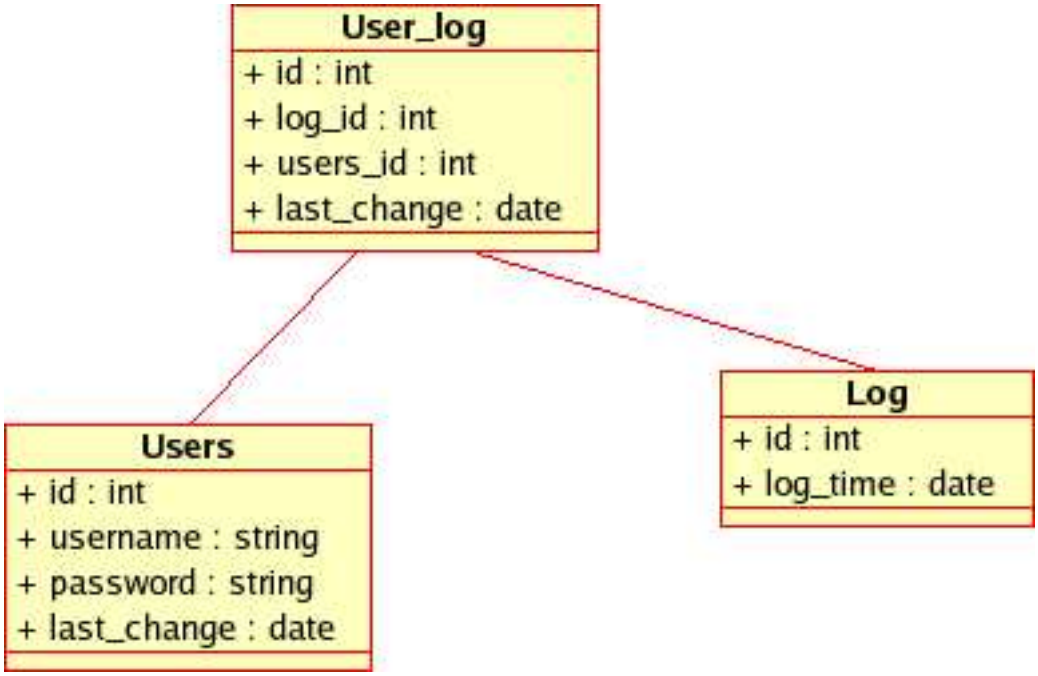
Atribuudid

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 2.3: Kasutajate autentimiseks ning andmete lisamis- ja muutmisaja 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.

2.4.8 Tabel Log

Log klass on mõeldud kasutajate logimise haldamiseks. Eesmärgiks on hoida unikaalseid identifikaatoreid, et oleks võimalik siduda andmete lisamine, muutmine või kustutamine selle teostamise ajaga.

Atribuudid

Log_time, mis salvestab sisse logimise aja

2.4.9 Tabel User

On eeskätt kasutajate tuvastamiseks ning andmete lisajate kirjeldamiseks loodud klass. Igale uuele andmebaasi kasutajale antakse kasutajanimi ja parool, et oleks võimalik üheselt identifitseerida andmete lisajat, muutjat või kustutajat.

Atribuudid

username hoiab kasutajanimesid

password on vajalik kasutajate üheseks turvaliseks tuvastamiseks

2.4.10 Tabel User_log

Antud klass võimaldab siduda logimisajad kasutajanimedega ja seeläbi on üheselt leitav nii andmete lisaja kui lisamisaeg. Tabeli identifikaatorvälja väärtus lisatakse kõigile bioloogilist infot sisaldavatele tabelitele. User_log tabeli **id** väärtus võimaldab soovi korral kustutada kõik andmebaasi konkreetse kasutaja poolt kindlal ajahetkel lisatud andmed kui selgub, et andmete lisaja on tol hetkel teinud vigu.

Atribuudid

log_id on klassi Log identifikaator

user_id on klassi User identifikaator

2.5 Kasutusjuhud

Kasutuslugu ehk kasutusjuht (*Use Case*) on järjekord toimingutest ja seostest kirjeldatava süsteemi ning selle kasutaja (*Actor*) vahel. Neid kasutatakse peamiselt süsteemi funktsionaalsete võimaluste väljendamise vahendina. Kogu süsteemi funktsionaalsus määratletakse kasutusjuhtude komplektiga, kus iga kasutuslugu esindab spetsiifilist sündmuste voogu. Kasutuslugu 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 2003).

Järgnevalt esitatakse bioloogilisi kasutuslugusid, mille abil on võimalik seletada **BiGeR** andmebaasi funktsionaalust ning antakse ülevaade andmebaasis realiseeritud päringutega. Allpool toodud kasutuslood illustreerivad andmebaasi tööd. Toodud päringud on aluseks graafiliste kasutajaliideste loomisel ning sellised päringud toimuvad andmebaasist veebiliidese kasutamisel.

2.5.1 Konkreetne transkriptsioonisait kindla geeni ees

Kasutajal on andmed transkriptsioonifaktori kohta ning selle esinemissait konkreetse geeni ees. Eksisteerivad seosed transkriptsioonifaktori ja geeni vahel, samuti transkriptsioonifaktori ja seondumissaidi vahel ning geeni ja seondumissaidi vahel. Vaja on arvestada kolme seosega:

- 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 |
+------+---------+---------+-----------+
```

2.5.2 Transkriptsioonifaktori konserveerunud sekvents ja loetelu geenidest, mille järgi see on 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 ja 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 |
+----+------+
```

2.5.3 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 geeni SPR3 transkriptsioonialgussaidid (TSS) ja nende 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.

2.5.4 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. Ek-

sisteerivad 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.

2.5.5 Klasterdamisel saadud *in silico* saidi kirjeldused

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

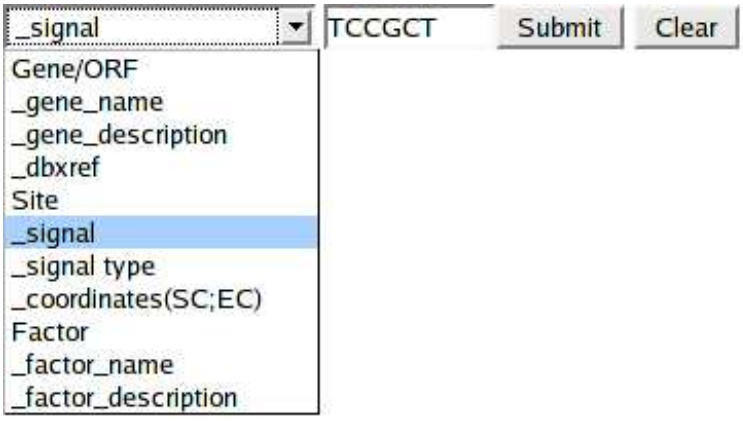
- seondumissait on geeni ülesvoolu järjestuses

Andmete lisamine Esimeses etapis tuleb kirjeldada **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**.

2.6 BiGeR-i veebiliides

Andmebaasi kergemaks kasutamiseks ning päringute sooritamiseks töötati välja veebiliidese esialgne prototüüp. Kasutajaliides ning päringusüsteem on kirjutatud keeltes Perl ja HTML. Veebipõhiselt on võimalik teostada päringuid kasutades spetsiaalset vormi ning täpsustades oma päringuid rippmenüüst valitava tabeli abil (joonis 2.4). Päringule vastavad andmed kuvatakse

veebilehele HTML formaadis ning samuti kirjutatakse päringule spetsiiflised väärtused tekstifaili.



Joonis 2.4: **BiGeR**-i päringuvorm ja rippmenüü

BiGeR-i kasutajaliidese selgemaks mõistmiseks on allpool toodud näidispäring. Päringu eesmärgiks oli küsida andmebaasist kõik sellised seondumissaidid, mille alamstringiks oleks järjestus `TCCGCT`. Esmalt tuleb valida rippmenüüst tabeli *Site* väli *_signal* ning sisestada tekstiväljale otsitava alamstring `TCCGCT` (joonis 2.5 punkt 1). Tulemuseks on veebileht (joonis 2.5 punkt 2), mille päises on link tekstifailile (joonis 2.5 punkt 3). Veebilehel kuvatakse leitud alamstringid ning neid kirjeldav informatsioon. Selgitava info ülesehitus on järgnev: `Ac` on *accession number*, mis viitab andmebaasi kirjele kust andmed pärit. Käesoleval juhul `R01043` on viide `TRANSFAC`-i (Wingender *et al.* 2000) andmebaasi. `Coordinate`s on transkriptsioonifaktori seostumissaidi koordinaadid arvestades geeni alguses, mille ees nad esinevad. Joonisel 2.5 geeniks CUP1 ning koordinaatideks -112;-146. Selgitavas infos on ka viited esialgsetele artiklitele `Medline`-i andmebaasi (Medline 2004).

2.7 Andmebaasi statistika

Seisuga 23.05.2004 on andmebaasis andmeid:

- neljast eri allikast, seejuures kahest olemasolevast andmebaasist (Wingender *et al.* 2000; Zhu & Zhang 1999) ning lisaks kahes artiklis (Kellis *et al.* 2003; Lee *et al.* 2002) avaldatud andmed.
- 606 geeni
- 232 faktorit
- 1291 saiti, nendest:
  - 1057 oligot
  - 0 konsensusjärjestust
  - 195 transkriptsiooni algussaiti
  - 39 maatriksit
- 1317 geeniregulatsiooni kirjet, millest
  - 639 kirjet sisaldavad infot nii transkriptsioonifaktori, geeni kui seondumissaidi kohta
  - 98 kirjet transkriptsioonifaktorite ja geenide omavaheliste seoste kohta
  - 39 kirjet geenide ja saitide kohta (TSS-id)
  - 9 kirjet seondumissaidi ja transkriptsioonifaktori kohta

Andmebaasi loomiseks ja andmete töötlemiseks on kirjutatud kaksteist andmetöötlus- , päringu - ja andmete sisestamise programmi, kogumahus 1800 rida.

1.

BiGeR has currently:

Genes: 838, Sites: 1291, Signals: 1291, Regulation events: 1291

_signal | TCCGCT | Submit | Clear

2.

your result as txt file

Regulated by factor

Ac is: R01043 coordinates are: (-112;-146) description is: CUP1 (metallothionein); Gene: G000902. |
of the yeast metallothionein gene Mol. Cell. Biol. 11:1232-1238 (1991). [2] MEDLINE; 89098931. F
upstream activating sequences Proc. Natl. Acad. Sci. USA 86:65-69 (1989). site is:

CGATGCGTCTTTTCCGCTGAACCGTT.

and type is: oligo

Regulated by factor

Ac is: R01846 coordinates are: (-118;-153) description is: CUP1 (metallothionein); Gene: G000902. |
change in CUP2 alters its mode of DNA binding Mol. Cell. Biol. 10:4778-4787 (1990). [2] MEDLIN
between Cu(I) and yeast ACE1 protein Proc. Natl. Acad. Sci. USA 86:5267-5271 (1989). [3] MEDL
altering the conformation of a specific DNA binding protein Cell 55:705-717 (1988). site is:

gatGCGTCTTTTCCGCTGAACCGttc.

and type is: oligo

3.

```
>158 oligo      R01043
 CGATGCGTCTTTTCCGCTGAACCGTT

>209 oligo      R01846
 gatGCGTCTTTTCCGCTGAACCGttc

>858 oligo
GCGTCTTTTCCGCTGAACCG
>862 oligo
GATGCGTCTTTTCCGCTGAACCGTTCCAGCAAAAAAGACTA
```

Joonis 2.5: Näidispäring andmebaasist **BiGeR** 1. Vormi täitmine ning tabeli ja välja valik. 2. HTML formaadis tulemus koos viidetega teistele andmebaasidele. 3. Tekstifail leitud oligonukleotiididega.

Arutelu

Käesoleva töö raames on valminud spetsiaalselt geeniregulatsiooni kirjeldamiseks mõeldud andmebaas **BiGeR**. Andmebaasi eesmärgiks on võimaldada olemasolevate teadmiste ühendamist geeniregulatsioonist, eriti transkriptsiooni kontrollist, ja uute ennustuste kirjeldamist. Tähtsal kohal on ka valmisolek eksperimentaalsete ja ennustuslike meetodite poolt saadud tulemuste esitamiseks ja võrdlemiseks. Andmebaasi olulise omadusena tuleb välja tuua spetsiaalselt geene, transkriptsioonifaktoreid ja transkriptsioonifaktorite seondumissaite ja nende seoseid ühendava tabeli *Regulation* olemasolu. Käesolev tabel on keskne edaspidisteks geeniregulatsioonivõrgustike modelleerimisteks.

Andmed pärinevad erinevatest allikatest ja nende ülesehitus ning geeniregulatsiooni kirjeldava info sisaldus on erinev. Kaasatud on nii transkriptsioonifaktoreid, nende seondumissaite kui reguleeritavaid geene kirjeldavad andmehulgad kahest varasemast andmebaasist SCPD(Zhu & Zhang 1999) ja TRANSFAC(Wingender *et al.* 2000). Näitena võib tuua **Site** tabeli, milles on kirjeldatud 1291 kirjet, millest 571 pärinevad TRANSFAC-i andmebaasist (Wingender *et al.* 2000), 648 on pärit SCPD(Zhu & Zhang 1999) andmebaasist ja 72 on pärit (Kellis *et al.* 2003) artiklist.

BiGeR-i eelis varem loodud transkriptsioonifaktorite kirjeldusi sisaldavate andmebaaside ees on peamiselt suunatus geeniregulatsiooni modelleerimisele ning avatus ka mitte eksperimentaalsete andmete kogumiseks. Töö käigus on püütud leida parim transkriptsioonifaktorite ja geenide omavahelisi seoseid kirjeldav andmestruktuur.

Kuigi andmebaasi esimeses etapis on kaasatud vaid *S. cerevisiae* andmed, on andmebaasi arenguks olulised ka kõrgematele eukarüootidele spetsiifiliste andmete kirjeldamine. Kõrgemate eukarüootide nagu inimese, hiire ning teiste imetajate andmed vajavad lisaks olemasolevatele võimalustele ka alternatiivsplaissingu kirjeldamist. Samuti eristavad kõrgemaid eukarüoote seni **BiGeR**-is käsitletud *S. cerevisiae*-st pikemad promooterregioonid, võimendajate ja vaigistajate olemasolu ning võimalikud geeniregulatsiooni mõjutavad signaalid intronites. Nende spetsiifiliste omaduste bioinformaatiline toetus on plaanitud andmebaasi **BiGeR** järgmiseks arenguetapiks.

Kokkuvõte

Viimastel aastatel toimuv suuremahuline geeniregulatsiooni mehhanismide eksperimentaalne uurimine vajab spetsiaalselt selleks otstarbeks loodud andmebaaside toetust. Lisaks on vajalikud andmebaasidega integreeritud *in silico* meetodid, mis võimaldavad geeniregulatsiooni modelleerimist.

Antud töö teoreetilises osas anti kirjanduse ülevaade geeniregulatsiooni mehhanismidest ning pikemalt käsitleti transkriptsiooni ja selle kontrolli. Samuti kirjeldati bioloogiliste andmebaaside peamisi omadusi. Enamlevinud meetoditest kirjeldati *in vitro* ja *in silico* eksperimente transkriptsioonifaktorite seondumissaitide määramiseks. Töös anti lisaks ülevaade transkriptsioonifaktorite seondumissaitide erinevatest esitusviisidest ning võrreldi nende häid ja halbu külgi.

Praktilises pooles anti ülevaade uuest geeniregulatsiooni andmebaasist **BiGeR**, mis võimaldab integreerida erinevates juba eksisteerivates andmebaasides olevad andmed üheks tervikuks ja samas toetab uute andmete sisestamist ning analüüsi. Andmebaasi modelleerimiseks õpiti tundma geeniregulatsioonis osalevate bioloogiliste olemite omavahelisi seoseid ning neist lähtuvalt kujundati andmebaasi struktuur. Loodi ka meetodid andmete töötlemiseks, millega erinevatest allikatest pärinevad andmed viiakse ühtsele kujule ning vastavusse meie poolt välja töötatud andmestruktuurile. Peale andmetöötlust on erinevatest allikatest pärit andmed omavahel võrreldavad ning ühildatavad.

BiGeR on kasutatav ka teiste organismide geeniregulatsiooni andmete kirjeldamiseks ja modelleerimiseks. Projekti edasine eesmärk on arendada ja lisada tööriistu nii arvutiprogrammide kui ka tavakasutajate jaoks. 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 the presence of many complete DNA sequences, regulatory signals can be studied in the DNA. The aim of our work is to create a database for storing and analyzing data about gene regulation, and to facilitate further analysis *in silico*.

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.

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 the control mechanisms of gene regulatory and mainly we introduce transcription regulation. 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, matrices, consensus sequences and regular expressions. Also, we describe 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. Also overview of the first web interface prototype is given.

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).

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* 723–750.

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

Blackwood, E., and Kadonaga, J. 1998. Going the distance: a current view of enhancer action. *Science*.

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.

Buratowski, S. 1996. Transcription Factor IIB; http://tfiib.med.harvard.edu/transcription/tfiib.html.

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

Charniak, E. 1991. Bayesian networks without tears. *Artificial Language 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. 2003. Use Case Alternate Intro; http://members.aol.com/acockburn/papers/altintro.htm.

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.

Elmasri, R., and Navathe, S. B. 2000. *Fundamentals of Database Systems.* Addison Wesley Longman, Inc, Third edition.

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

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).

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

Kimball. 2001. Gene regulation in eukaryotes; http://users.rcn.com/jkimball.ma.ultranet/biologypages/p/promoter.html.

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.

Lane, D.; Prentki, P.; and Chandler, M. 1992. Use of gel retardation to analyze protein-nucleic acid interactions. *Microbiol Rev.* 56:509–528.

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

Lee, T. I.; Rinaldi, N. J.; Robert, F.; Odom, D. T.; Bar-Joseph, Z.; Gerber, G. K.; Hannett, N. M.; Harbison, C. T.; Thompson, C. M.; Simon, I.; Zeitlinger, J.; Jennings, E. G.; Murray, H. L.; Gordon, D. B.; Ren, B.; Wyrick, J. J.; Tagne, J.-B.; Volkert, T. L.; Fraenkel, E.; Gifford, D. K.; and Young, R. A. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*.

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.

Maimets, T. 1999. *Molekulaarne rakubioloogia*. Ilmamaa.

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.

Mallery, C. 2004. Gene regulation control in Eukaryotes; http://cats.med.uvm.edu/cats_teachingmod/microbiology/courses/gene_regulation/euk_regulation/4.1.grg.hub.intro.html.

Medline. 2004. Entrez PubMed; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.

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.

Nucleic Acid Research. 2004. Protein-nucleic acid interaction; http://nar.oupjournals.org/cgi/collection/protnucacidint.

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.

Peterson, H. 2004. Fülogeneetilise jalajälje meetod regulatoorsete järjestuste leidmiseks. In Vilo, J., ed., *Bioinformaatika II*, I. Tartu Ülikool, TÜMRI.

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.; 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.

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.