TARTU ÜLIKOOL
BIOLOOGIA-GEOGRAAFIATEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

EERO RAUDSEPP

BIOLOOGILISTE RADADE ANDMEBAASID

Bakalaureusetöö

Juhendaja: Jaak Vilo, PhD

Tartu 2005

SISUKORD

Lühendid ja mõisted

ATP – adenosiintrifosfaat	Adenosine Triphosphate
BIND – Biomolekulaarsete Interaktsioonivõrgustike Andmebaas	Biomolecular Interaction Network Database
DBMS – andmebaasi juhtimissüsteem	Database Management System
DIP – Interakteeruvate Valkude Andmebaas	The Database of Interacting Proteins
DNA – desoksüribonukleiinhape	Deoxyribonucleic Acid
EBI – Euroopa Bioinformaatika Instituut	European Bioinformatics Institute
EC – Ensüümi Komisjon	Enzyme Commission
EMP – Ensüümide ja Metaboolsete Radade Andmebaas	Enzymes and Metabolic Pathways Database
GO – Geeniontoloogia	Gene Ontology
GUI – graafiline kasutajaliides	Graphical User Interface
HTML – hüperteksti märgistuskeel	HyperText Markup Language
KEGG – Kyoto Geenide ja Genoomide Entsüklopeedia	Kyoto Encyclopedia of Genes and Genomes
KGML – KEGGi märgistuskeel	KEGG Markup Language
MAPK – mitogeen-aktiveeritud valgu kinaas	Mitogen-Activated Protein Kinase
MNV – Metaboolse Võrgustiku Visualiseerija	Metabolic Network Visualizer
MIPS – Müncheni Valgu Järjestuste Infokeskus	Munich Information Center for Protein Sequences
MOD – mudelorganismi andmebaas	Model Organism Database
mRNA – matriits-RNA	Messenger-RNA
NCBI – Riiklik Biotehnoloogia Infokeskus	National Center for Biotechnology Information
OMIM – Inimese Pärilike Haiguste Andmebaas	Online Mendelian Inheritance in Man
ORF - avatud lugemisraam	Open Reading Frame
PGDB – radade/genoomi andmebaas	Pathway/Genome Database
RNA – ribonukleiinhape	ribonucleic acid
SBML – süsteemibioloogia märgistuskeel	System Biology Markup Language
SQL – struktureeritud päringukeel	Structured Query Language
UML – ühtsustatud modelleerimiskeel	Unified Modeling Language
VRML – virtuaalse reaalsuse märgistuskeel	Virtual Reality Markup Language

Sissejuhatus

Ajalooliselt algas molekulaarsete võrgustike arvutuslik analüüs bakterite biokeemiliste radade kirjeldamisega. Enne seda olid uuringud suunatud üksikute geenide ja geeniperekondade struktuuri ja funktsiooni selgitamisele. Kaasaegses teaduses on peamine probleem raku biokeemilise võrgu ehituse mõistmine. Püütakse ennustada antud geenide põhjal valgu interaktsioonivõrgustikke, mis vastutavad mingi kindla protsessi eest rakus. Suureks väljakutseks on kogu raku ja organismi protsesside kujutamine arvutis.

Paljud olemasolevad molekulaarbioloogia andmebaasid on koondanud tähelepanu just bioloogiliste makromolekulide (DNA, RNA, valgud), aga ka genoomide järjestustele ja struktuursetele omadustele. Tänapäeval on üha enam hakatud uurima interaktsioonimehhanisme geenide ja valkude vahel. Neid interaktsioone võib vaadelda kui võrgustikku rakulistest protsessidest, mis sisaldavad metaboolseid radu, signaali ülekande süsteeme ja regulatoorseid võrgustikke. Kirjeldamaks metaboolseid radu, reaktsioone, ensüüme ja metaboliite, on loodud mitmeid andmebaase, mis sisaldavad andmete paremaks esituseks ja töötluseks visualiseerimis- ja analüüsitööriistu.

Käesoleva töö eesmärk on luua andmebaas biokeemilistest radadest ja nendega seotud geenidest. Kuna vead bioloogilistes radades põhjustavad mitmeid erinevaid haigusi, siis on andmebaasi haaratud ka vastavad haigused kommentaaride ning viidetega. Lisaks on andmebaasis kirjeldatud hulgaliselt ensüüme, keemilisi ühendeid ja reaktsioone, mis on ühtlasi bioloogiliste radade alustalaks.

Töö esimeses pooles kirjeldatakse enam levinud metaboolsete radade ning valk-valk interaktsioonide andmebaase. Vaatluse all on nende sisu, mudelid, autorite poolt loodud tööriistad ning eesmärgid. Samuti sisaldab töö teoreetiline osa kokkuvõtvat peatükki metaboolsete radade visualiseerimisest. Töö teises pooles antakse ülevaade autori poolt koostatud bioloogiliste radade andmebaasist. Põhjalikult on lahti seletatud andmebaasi mudel oma atribuutidega. Lisaks on esitatud näidispäringud, nende tulemused ja andmebaasile loodud veebiliidese tutvustus. Praktilise osa lõpuosas on näidatud ka andmebaasi statistilised andmed.

I Kirjanduse ülevaade

1. Radade andmebaaside eesmärgid ja väljakutsed

Rada on biokeemiliste reaktsioonide ahel, kus ühe reaktsiooni produkt on substraadiks või ensüümiks järgnevale reaktsioonile (Karp, 2001). Radade andmebaas kirjeldab biokeemilisi radu, reaktsioone ja ensüüme. Enim kasutatavad radade andmebaasid sisaldavad lisaks metaboolsetele radadele ka signaali ülekande ja regulatoorseid radasid. Seega, kui genoomi andmebaasidest saab informatsiooni sekveneeritud organismidest, siis radade/genoomi andmebaas ühendab tervikuks radade ja organismi kogu genoomi informatsiooni.

Metaboolsete radade konstrueerimise võib jagada neljaks etapiks (Karp, 2001):

1. avatud lugemisraamide (ORF) identifitseerimine
2. ensüümi kodeerivate geenide ennustamine
3. EC (Enzyme Commission) numbrite määramine
4. geenide ja ensüümide kaardistamine teadaolevatesse radadesse.

Esimese kahe sammu jaoks on vaja järjestuste andmebaase, kolmanda sammu jaoks ensüümi nomenklatuuri andmebaasi ja seejärel teadaolevate metaboolsete radade andmebaasi. Paraku ei saa kõiki geene radadesse kaardistada, põhjuseks on sekundaarne metabolism või ebatavalistel tingimustel sisselülituv ainevahetus.

Radade andmebaaside koostamisel on mitmeid eesmärke, kuid eelkõige on nad loodud teadlastele päringute tegemiseks ja mustrite otsimiseks. Ühtlasi kergendavad radade andmebaasid genoomi andmete analüüsi. Tegu on kiirelt areneva bioinformaatika haruga, mistõttu nõudmised radade andmebaasidele kasvavad ja esitatud on mitmeid väljakutseid (Karp, 2001):

1. Suurte signaalivõrgustike modelleerimine organismides.
2. Eukarüootsetes organismides esinevate mahukate radade võrgustike skeemide automaatne koostamine.
3. Andmete kasutamine erinevates andmebaasides ja rakendusprogrammides.
4. Uute analüüsialgoritmide loomine haigusi põhjustavate radade võrgustiku analüüsiks, mis kaudselt aitab kaasa ravimitööstuse arengule.

Radade andmebaaside jaoks on loodud mitmeid algoritme, mis ennustavad metaboolseid produkte, mida organism võib antud kasvukeskkonnas toota (Karp, 2001). Kuna kõne all olevate andmebaaside maht on suur, tuleb disainida täpsed meetodid andmete elektrooniliseks haldamiseks. Samuti peavad biomolekulaarsete interaktsioonide ja biokeemiliste radade andmed vastama kindlatele nõuetele (Bader & Hogue, 2000):

1. Andmed peavad olema detailselt kirjeldatud.
2. Andmed peavad olema kergesti esitatavad arvutis.
3. Andmebaas peab olema inimestele kergesti arusaadav.

Järgnevalt tulevadki vaatluse alla levinumad nimetatud nõuetele vastavad valk-valk interaktsioonide ja bioloogiliste radade andmebaasid. Kuna valk-valk interaktsioonide andmebaaside hulk on väiksem võrreldes radade andmebaasidega, siis nendest kirjeldame lühemalt kolme levinumat. Bioloogiliste radade andmebaaside valikul on püütud samuti kirjeldada levinumaid, ent üksteisest erinevaid projekte. Kuna radade andmebaase on mitmeid, sai valik tehtud vastavalt andmebaaside mahtudele ning artiklitele, mis on vastavate andmebaaside kohta ilmunud. Antud töö lisades 1 ja 2 on vaatluse all olevatest andmebaasidest tehtud kokkuvõtvad tabelid, mis on toetavaks materjaliks lugemisele. Lisa 1 võtab kokku levinumate radade andmebaaside autorid, eesmärgid ja mahud. Lisa 2 toob välja baaside mudelid, eelised ning tööriistad.

Veel enne valk-valk interaktsioonide ja metaboolsete radade andmebaaside kirjeldamist on antud ülevaade Gene Ontology (GO) konsortsiumist, mis tegeleb geenide ja valkude ühtse terminoloogia väljatöötamisega ja millele paljud radade andmebaasid viitavad. Üha enam on hakanud andmebaaside autorid kirjeldama geene ja valke GO terminitega ja kuna radade andmebaasid pole erandiks, siis on järgnev peatükk vajalik taustinformatsiooniks.

2. Gene Ontology konsortsium

Geeni ontoloogia konsortsium (GO Consortium) loodi 1998.a. kolme mudelorganismi andmebaasi vahel: FlyBase (The FlyBase Consortium, 2003), Mouse Genome Informatics (MGI) (Blake *et al.*, 2003) ja *Saccharomyces* Genome Database (SGD) (Dwight et al., 2004). 2000.a. lisandusid The *Arabidopsis* Information Resource (TAIR) (Rhee et al., 2003) ja *Caenorhabditis elegans* grupid. GO eesmärk on luua ühine, struktureeritud, dünaamiline, täpselt defineeritud sõnastik geenide ja geeniproduktide kirjeldamiseks suvalises organismis.

Seejuures arvestatakse, et teadmised geenide ja valkude funktsioonide kohta pidevalt suurenevad. Kõik nimetatud grupid annoteerivad geene ja geeniprodukte GO sõnastiku terminitega. Lisaks pakub GO tööriistu sõnastikust päringute tegemiseks.

GO ei ole ette kirjutatud standard, mistõttu ta ei määra nomenklatuuri andmebaaside vahel. Samuti ei määratle GO homoloogiaid erinevate organismide geeniproduktide vahel. GO tugevus seisneb bioloogiliste sõnastike spetsiifilisuses ning täpsete suhete loomises terminite vahel. GO-s konstrueeritakse kolm iseseisvat ontoloogiat:

1. Bioloogiline protsess - sündmus, millele aitab kaasa geen või geeniprodukt, näiteks raku kasv, signaali ülekanne, translatsioon.
2. Molekulaarne funktsioon - geeniprodukti biokeemiline aktiivsus, näiteks ensüüm, transporter, ligand.
3. Rakuline komponent - asukoht rakus, kus geeniprodukt on aktiivne, näiteks ribosoom, proteosoom jt. (The Gene Ontology Consortium, 2000; The Gene Ontology Consortium, 2001).

GO terminite defineerimisel kasutatakse võimalikult palju raamatut „The Oxford Dictionary of Molecular Biology“ (1997), samuti SWISS-PROT (Boeckmann *et al.*, 2003) andmebaasi. GO terminid ei ole liigispetsiifilised, kuid kehtivad klassi tasemel ning on varustatud vastavate viidetega. Ontoloogiad on salvestatud tekstifaili formaati (*flat-file*). GO-sse lisatakse termineid ainult projektis osalevate andmebaaside haldajate poolt.

3. Valk-valk interaktsioonid

Iga rakuline protsess on reguleeritud valk-valk interaktsioonide poolt, näiteks ensüümide fosforüleerimine kontrollib signaali ülekande kaskaade ja keeruliste komplekside tööd. Rakus hinnatakse toimuvat 2-10 valk-valk interaktsiooni ühe valgu kohta (Bader *et al*., 2001). Samas raku alamüksuses asuvad valgud interakteeruvad üksteisega tihedamini, kui erinevates raku osades paiknevad valgud. Samuti interakteeruvad valgud üksteisega erineva afiinsusega[1] ja erineval ajal. Interaktsioonide detekteerimine ja kirjeldamine on tihti raskendatud, sest paljud interaktsioonid on lühiajalised ja nõrgad. Seetõttu on vähesed interaktsioonid kirjeldatud energia ja kineetika terminitega. Tänu erinevate eksperimentaalsete meetodite rakendamisele on viimastel aastatel tuvastatud interaktsioonide hulk suurenenud eksponentsiaalselt. Nimetatud meetoditest on kolm tähtsamat pärmi kaksikhübriidi test (Ito *et al.*, 2000), valgu

[1] afiinsus - molekulide vahelise vastastikuse toime tugevus

kiibid (Zhu *et al.*, 2001) ning mass-spektraalanalüüs (Ho *et al.*, 2002). Pärmi kaksik-hübriidi testi puhul detekteeritakse valk-valk interaktsioone otsese vaatlusega. Mass-spektraalanalüüsi kasutatakse rohkem multimeersete valkude tuvastamisel, et leida, millised valgud esinevad kompleksis kindlal ajahetkel. Lisaks kasutatakse röntgenkiirte kristallograafiat aatomi tasemel vaatluseks.

Valk-valk interaktsioonid jagunevad oma olemuselt kolme funktsionaalsesse kategooriasse (Xenarios & Eisenberg, 2001):

1. metaboolsed ja signaali rajad, näiteks tsitraaditsükkel, MAPK (Mitogen-Activated Protein Kinase) signaali rada
2. morfogeensed rajad, näiteks organogenees, somaatiline embrüogenees.
3. struktuursed kompleksid, näiteks proteasoom, holoensüüm.

Valk-valk interaktsioonide andmebaasid sisaldavad kõiki nimetatud kategooriaid. Näited erinevatest valk-valk interaktsioonide andmebaasidest:

DIP (Database of Interacting Proteins) – eksperimentaalselt määratud interaktsioonid, paariviisilised interaktsioonid valkude vahel. Interaktsiooni kvaliteeti illustreerib visualiseerimisel joone paksus (Xenarios *et al.*, 2002).

BIND (Biomolecular Interaction Network Database) – valk-valk, valk-RNA, valk-DNA interaktsioonid molekulaarsest kuni radade tasemeni (Bader *et al.*, 2001).

IntAct – EBI (European Bioinformatics Institute) valk-valk interaktsioonide andmebaas, mis sisaldab ca 2200 interaktsiooni. Andmed pärinevad kirjandusest ja andmebaasi haldajatega teevad koostööd SWISS-PROT autorid (Hermjakob et al., 2004).

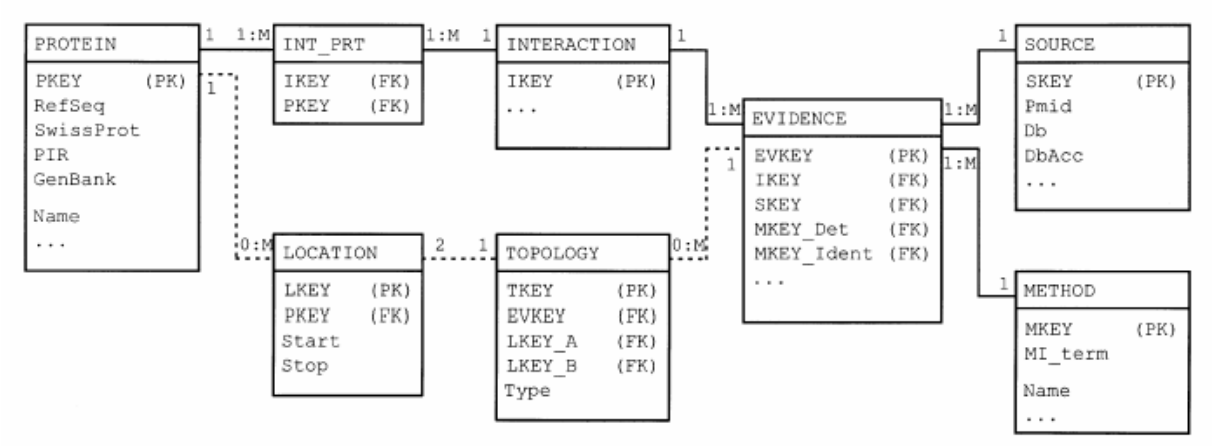
MIPS (Munich Information Center for Protein Sequences) – valgu aminohappelise järjestuse informatsioon pärmi ja teiste mudelorganismide kohta (Mewes *et al.*, 2000).

PROTEOME – valkude interaktsioonid, lokalisatsioonid, rakulised funktsioonid. Organismid: pärm, ussid, inimene (Costanzo *et al.*, 2001).

Kuna nimetatud andmebaasidest esimesed kaks (DIP ja BIND) on väga laialt kasutusel ning kolmas (IntAct) kiirelt arenev, siis peatume nendel veidi lähemalt.

3.1 DIP (Database of Interacting Proteins)

DIP relatsiooniline andmebaas (http://dip.doe-mbi.ucla.edu) sisaldab enam kui 18500 interaktsiooni rohkem kui 80 organismist (peamiselt *Saccharomyces cerevisiae*, *Helicobacter pylori* ja *Homo sapiens*). DIP kasutab PostgreSQL andmebaasi juhtimissüsteemi (http://www.postgresql.org). Andmebaasi lihtsustatud mudel on kujutatud joonisel 1.



Joonis 1: DIP andmebaasi mudel

Mudeli peatabeliteks on PROTEIN, SOURCE ja EVIDENCE, mis sisaldavad informatsiooni vastavalt üksikute valkude, eksperimentaalse info allika ja eksperimendi kohta. Info valk-valk interaktsioonide kohta on tabelites INTERACTION ja INT_PRT, et oleks võimalik kirjeldada binaarseid interaktsioone ja multi-valkude komplekse. TOPOLOGY tabel kirjeldab valgu komplekside kuju. LOCATION tabel sisaldab teavet interaktsioonis osalevate valkude asukohast (Salwinski *et al.*, 2004).

DIP-s kasutatakse informatsiooni eelkõige nendest teadusartiklitest, kus kirjeldatakse eksperimentaalselt tuvastatud interaktsioone. Andmebaasi kiire kasvu kaudseteks põhjusteks on valk-valk interaktsioonidega seotud artiklite arvu suurem publitseerimine ja erinevate eksperimentaalsete meetodite laiem kasutamine. Lisaks on informatsiooni hulk suurenenud PDB-s (Protein Data Bank) (Westbrook *et al.* 2002) olevate valgukomplekside struktuuride analüüsimise tulemusena. Artiklitest pärinevat infot saavad sisestada ainult andmebaasi haldajad. Loodud on viited teistele andmebaasidele nagu SWISS-PROT, TRANSPATH (Krull *et al.*, 2003), KEGG (Kanehisa *et al.*, 2004), YPD (Hodges *et al.*, 1999). Visualiseerimiseks on loodud Java-põhine tööriist JDIP. Tulevikus on andmebaasi haldajatel kaks peaeesmärki: suurendada inimese valkude osakaalu ning luua lisatööriistu andmete analüüsiks (Xenarios *et al.*, 2002).

3.2 BIND (Biomolecular Interaction Network Database)

BIND (http://binddb.org) on loodud kasutades objekt-relatsioonilist mudelit. BIND-i objektiks võib olla valk, DNA, RNA, ligand[2], molekulaarne kompleks või interaktsioon. Iga objekti kohta on salvestatud tema nimi, sünonüümid, päritolu (looduslik või mitte), asukoht

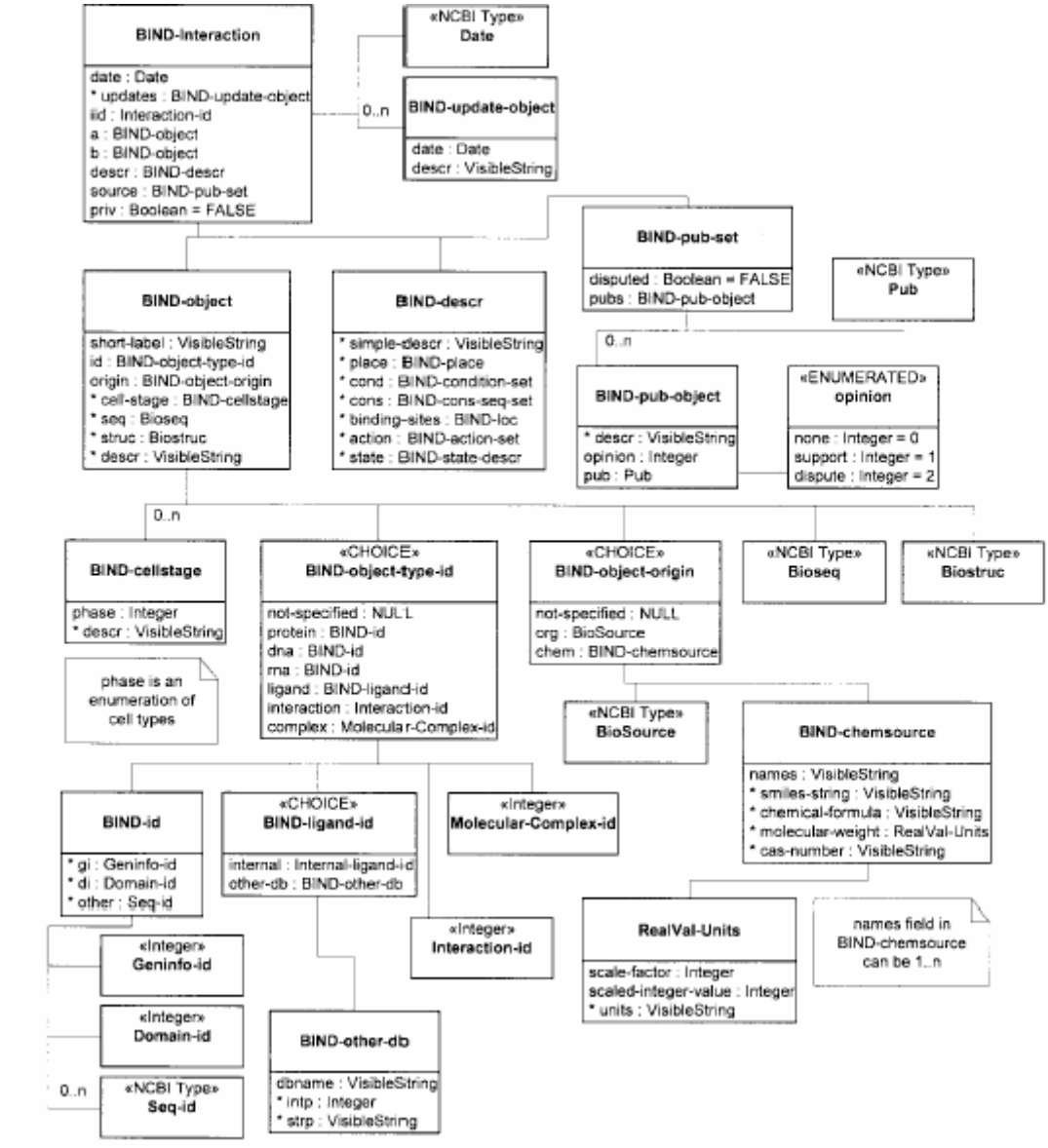
[2] ligand - molekul, mis seondub valgu või muu molekuli komplementaarse alaga

rakus, viited järjestuste andmebaasidesse ja 3D struktuur. Päritolu on oluline liikidevaheliste interaktsioonide kirjeldamiseks, valdavalt on need viiruse ja peremeesorganismi vahelised valk-valk interaktsioonid.

BIND-i 3 peamist andmetüüpi on:

1. Interaktsioon – kirjeldus molekulide A ja B seondumisest.
2. Molekulaarne kompleks – stabiilne molekulide kompleks, mis omab funktsiooni (näiteks ribosoom). Siin kirjeldatakse ka kompleksi topoloogiat ja järjestikusi interaktsioone, mille tulemusena kompleks moodustub.
3. Rada – grupp erinevaid molekule, mis moodustavad interaktsioonide võrgustiku.

BIND andmemudel on väga mahukas, kuid selle olulisem osa on kujutatud joonisel 2.



Joonis 2: BIND andmebaasi mudel

Mudel ei ole lõplik, sest BIND-interaction ja BIND-descr omavad alammudeleid. Samuti puuduvad jooniselt objektid BIND-Molecular-Complex ja BIND-Pathway.

BIND-object tabel kirjeldab keemilisi objekte (aatomid, molekulid või molekulide kompleksid) ja sisaldab järgmisi atribuute: *short-label*, *BIND-object-type-id*, *BIND-object-origin, BIND-cellstage, NCBI Bioseq, NCBI Biostruc* ja *descr*.

BIND-descr tabel kirjeldab interaktsioone kahe molekuli või aatomi vahel. BIND-interaction atribuudid: *simple-descr, BIND-place, BIND-condition-set, BIND-cons-seq-set, BIND-loc, BIND-loc-gen, BIND-action-set* ja *BIND-state-descr*.

Kui rohkem kui kaks interaktsiooni moodustavad kompleksi, siis neid kirjeldab BIND-Molecular-Complex tabel, millel on BIND-interaction-ga sarnased atribuudid.

Kui rohkem kui kaks interaktsiooni moodustavad raja, siis neid kirjeldab BIND-Pathway tabel, kus on info metaboolsete ja signaali radade kohta.

BIND andmebaasis on detailselt kirjeldatud interaktsioonid, molekulaarsed kompleksid ja rajad. Interaktsioonid on kirjeldatud rakulise asukoha, interaktsiooni vaatlemiseks kasutatud eksperimentaalsete tingimuste, konserveerunud järjestuste, kineetika ja termodünaamika terminitega. Tänapäeval kasutusele võetud erinevate meetodite tõttu lisandub üha enam andmeid molekulaarsetest interaktsioonidest, radadest ja valkude posttranslatsioonilistest modifitseerimistest. Viimase kohta annab palju informatsiooni mass-spektromeetria. BIND-i eesmärgiks on olla andmekaevanduse ja interaktsiooni informatsiooni visualiseerimise platvorm (Bader *et al*., 2001).

3.3 IntAct

IntAct (http://www.ebi.ac.uk/intact) on objekt-relatsiooniline andmebaas, mis sisaldab hetkel ca 2200 valk-valk interaktsiooni. IntAct-i andmemudel koosneb kolmest põhikomponendist: Experiment, Interaction ja Interactor (eksperiment, interaktsioon ja interaktsioonis osaleja). Experiment kirjeldab eksperimentaalseid tingimusi, mida on kasutatud interaktsioonide tuvastamiseks. Interactor on bioloogiline üksus (tavaliselt valk, aga võib olla ka DNA järjestus), mis võtab interaktsioonist osa. Interaction sisaldab ühte või enamat interaktsioonist osa võtvat Interactor-it. IntAct-i andmemudel on mõjutatud nii DIP (Xenarios *et al*., 2002) kui BIND (Bader *et al*., 2001) mudelite poolt. IntAct viitab igal võimalikul juhul NCBI ja GO andmebaasidesse. IntAct-i lähim eesmärk on arendada välja süsteem, mis võimaldaks regulaarset andmete vahetust peamiste valk-valk interaktsioonide andmebaaside vahel (Hermjakob et al., 2004).

4. Ülevaade levinumatest radade andmebaasidest

Mahukamad ja enim kasutatavad radade andmebaasid on KEGG (Kanehisa *et al.*, 2004), EcoCyc (Karp, Riley, Saier *et al.*, 2002), MetaCyc (Karp, Riley, Paley *et al.*, 2002), aMAZE (Lemer *et al.*, 2004) ja Reactome (Joshi-Tope *et al.*, 2003). Nimetatutest võib levinuimaks nimetada KEGG-i. EcoCyc kasutab sama andmebaasiskeemi, mis MetaCyc. Kui EcoCyc on ainult *E. coli*-le orienteeritud andmebaas (signaali ülekande rajad, transporterid, geenid jne.), siis MetaCyc-is on erinevad liigid, sealhulgas inimene ja mõned imetajad. Põhirõhk on siiski suunatud mikroorganismidele. aMAZE projekt analüüsib peamiselt geeniregulatsiooni, biokeemilisi radu ja signaali ülekandeid. Reactome peaeesmärgid on suunatud eelkõige inimese bioloogiliste radade kirjeldamisele.

4.1 KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) projekt (http://www.genome.ad.jp/kegg) loodi 1995.a. ja selle eesmärk on esitada arvutis kõik molekulaarsed koostisosad ning molekulaarsete interaktsioonide võrgustik, et kirjeldada, kasutada ja ennustada elussüsteemide funktsionaalseid aspekte (Kanehisa *et al*., 2004). Koostisosade all peetakse silmas nii geene, geeniprodukte (DNA, RNA, valk) kui ka teisi keemilisi aineid elusrakkudes (näiteks metalli-ioonid). KEGG hõlmab hetkel informatsiooni 253 eri liigi bioloogiliste radade kohta. KEGG sisaldab kolme põhiandmebaasi:

1. PATHWAY ehk valkude võrgustik – molekulaarsete interaktsioonide võrgustike informatsioon (rajad ja kompleksid). PATHWAY andmebaas on veebis hierarhiliselt jagatud neljale tasemele: metabolism, geneetiline info, keskkonna info ja ülejäänud rakulised protsessid. Loodud on uus kategooria – inimhaigused. PATHWAY andmebaas on kolme tüüpi interaktsioonide või suhetega geeniproduktide võrgustik (Kanehisa *et al*., 2002):
    a) ensüüm-ensüüm suhted
    b) otsesed valk-valk interaktsioonid nagu seondumine ja fosforüleerimine
    c) transkriptsioonifaktorid ja sihtmärk-geeniproduktid.
2. GENES/SSDB/KO ehk Geeniuniversum – geenide ja valkude informatsioon, mis on loodud genoomi järjestuste põhjal. GENES andmebaas sisaldab infot enam kui 500000 geeni kohta, SSDB (Sequence Similarity Database) on järjestuse sarnasuste andmebaas ning KO (KEGG Orthology) sisaldab KEGG-i poolt loodud ortoloogiaid. Seega saab

Geeniuniversumist pärida näiteks ortoloog[3]/paraloog[4] suhteid, operoni[5] infot, suhteid geenide vahel täielikult sekveneeritud genoomides.

3. LIGAND ehk Keemiline universum – rakuliste protsessidega seotud keemiliste ühendite ja reaktsioonide informatsioon. LIGAND andmebaasi moodustavad neli andmebaasi:

a) COMPOUND – metaboliitide keemilised struktuurid jt. keemilised ühendid (ravimid, ksenobiootikumid[6])
b) GLYCAN – karbohüdraatide struktuurid
c) REACTION – keemilised reaktsioonid (enamjaolt ensümaatilised)
d) ENZYME – lisaväärtustega ensüümi nomenklatuuri andmebaas, mis sisaldab kolme osa (Kanehisa *et al*., 2004; Goto *et al*., 1998):
    1. ensüümi ja tema poolt katalüüsitava reaktsiooni kirjeldus
    2. ensüümiga seotud keemiliste ühendite kogum
    3. viited teistesse andmebaasidesse

Lisaks on KEGGis piiratud koguses eksperimentaalseid andmeid mikrokiibi geenieksпressiooni profiilide ja pärmi kaksik-hübriidsete süsteemide kohta salvestatud vastavalt EXPRESSION ja BRITE andmebaasidesse. KEGG omab viiteid mitmetele välisandmebaasidele, kuid ta on siiski iseseisev süsteem.

KEGG-s on genoom geenide graaf ning rada on geeniproduktide (enamasti valgud) graaf. Sobitades geene genoomis ja geeniprodukte radades, saab KEGG-i kasutada raku funktsioonidega seotud valkude interaktsioonivõrgustike ennustamiseks. KEGG-s ei ehitata iga liigi jaoks eraldi raja varianti, vaid luuakse üks suur raja skeem, kus on välja toodud liikidele omased iseärasused, kui need eksisteerivad. KEGG on täiendav allikas olemasolevatele järjestuste ja 3D-struktuuride andmebaasidele, suunates tähelepanu geenide ja valkude suhetele ning interaktsioonidele (Kanehisa *et al*., 2002). Lisaks metaboolsetele radadele plaanitakse arvutis esitada ka signaali ülekande ja rakutsükli geneetilisi radu.

Kui vanasti olid radade diagrammid KEGG-s saadaval ainult GIF või PNG failidena, siis nüüd on kõik metaboolsed ja mõned regulatoorsed rajad saadaval KGML-s (KEGG Markup Language), mis võimaldab KEGG-i radadega manipuleerimist (Kanehisa *et al*., 2004; Kanehisa, 1996).

[3] ortoloog – eri liikidest pärit järjestused, millel on üks ühine eellane
[4] paraloog – samast liigist pärit järjestused, millel on üks ühine eellane
[5] operon – bakterites sarnaste struktuursete geenide kogu, millelt sünteesitakse kokku üks mRNA
[6] ksenobiootikum – loodusvõõras aine

4.1.1 KEGG - LIGAND andmebaasi kirjeldus

KEGG-i Keemilise universumi andmebaasi LIGAND jaoks püstitati järgmised ülesanded: tundmatute geeniproduktide bioloogiliste funktsioonide identifitseerimine, radade konstrueerimine, erinevate liikide geenide ja genoomide võrdlev analüüs. LIGAND-i puhul kasutatakse tekstifaili (*flat-file*) formaati, et saaks ühenduda teiste andmebaasidega DBGET/LinkDB süsteemist. Siin tuleb *flat-file*-i käsitleda veidi teisiti – nimelt pole tegu tavalise tekstifailiga, vaid ta võib sisaldada pildifaile (GIF ja MOL formaadid) COMPOUND sektsiooni jaoks. LIGAND on ühtlasi KEGG-i ja DBGET/LinkDB süsteemide põhikomponent. Ta ühendab endas molekulide faktilised andmed (geenid ja geeniproduktid) ning bioloogilised suhted nende seas (molekulaarsed interaktsioonid ja rajad). Metaboolne rada konstrueeritakse täielikult sekveneeritud genoomi järjestuse info põhjal. LIGAND-ga on ühendatud EMP (Enzymes and Metabolic Pathways database) andmebaas, mis sisaldab radade kohta detailsemat informatsiooni (näiteks K_m[7] väärtused). Nii LIGAND-i kui EMP-d saab kasutada radade konstrueerimiseks, üleüldiseks analüüsiks kui ka infoallikana ravimi väljatöötamiseks (Goto *et al.*, 1998).

4.1.2 Binaarsed relatsioonid KEGG-s

KEGG-s kujutatakse geenide interaktsioone binaarse relatsioonina. KEGG-i teine tähtis kontseptsioon on hierarhia, mis kirjeldab geenide ja molekulide funktsionaalseid, struktuurseid ja evolutsioonilisi suhteid. Teisisõnu, binaarne relatsioon kujutab horisontaalseid, hierarhia vertikaalseid suhteid. Hierarhilisi klassifikatsioone peetakse binaarsete relatsioonide pikenduseks, mida saab kasutada ensüümi funktsioonide ennustamiseks. Radade arvutustes kasutatakse kahte tüüpi binaarseid relatsioone:

1. substraat-produkt relatsioon, mis on tuletatud LIGAND andmebaasist
2. ensüüm-ensüüm relatsioon, mis vastab ensüümide paarile raja diagrammides ehk teisisõnu on tegu lähimate naaberensüümidega.

KEGG-i kasutatakse tihti lühima tee arvutamiseks kahe komponendi vahel metaboolse raja diagrammis. Selle jaoks rakendatakse Dijkstra või Floydi algoritmi (Kanehisa, 1996; Ogata *et al.*, 1996). Dijkstra algoritm leiab lühima tee antud graafi punktist kõigi graafi punktideni. Floydi algoritm leiab lühima tee kahe graafi tipu vahel (Aho, Ullmann, 2000).

[7] K_m – Michaelise konstant, mis mõõdab ensüümi afiinsust substraadile (ensüümi katalüütilise aktiivsuse mõõt)

4.2 EcoCyc

E. coli on organismidest kõige detailsema geneetilise võrgustiku mudeliga. EcoCyc (http://ecocyc.org) on *E. coli* tüve K-12 andmebaas, mis sisaldab metaboolseid ja signaali ülekande radu, ensüüme, transportvalke ning geenieksprессiooni kontrollmehhanisme. EcoCyc on:

1. individuaalsete geenide tasemel mikroobsete genoomide analüüsiks
2. valk-valk interaktsioonide tuvastamiseks
3. radade evolutsiooni kirjeldamiseks.

EcoCyc kasutab objekt-orienteeritud andmemudelit, kus informatsioon on jagatud klassidesse sarnaste atribuutide alusel. Objekti klassideks on rajad, reaktsioonid, ensüümid, transporterid, geenid, promooterid, transkriptsiooniühikud jne. Transkriptsiooniühik on DNA regioon, mis sisaldab ühte promooterit, transkriptsioonifaktori seostumissaite, mis määravad transkriptsiooni initsiatsiooni sellelt promooterilt, ühte või mitut geeni, mis transkribeeritakse promooterilt ja transkriptsiooni terminaatorit. Transkriptsiooniühiku ja promooteri vahel on üks-ühele suhe.

Transkriptsiooniühiku ja operoni vahel on kaks peamist erinevust. Operonis võib olla mitu promooterit ja terminaatorit ning ta peab sisaldama rohkem kui ühte geeni. Transkriptsiooniühikus ei ole mitut promooterit ega terminaatorit ning ta võib sisaldada ühte või mitut geeni. Paraku on ainult 25% *E. coli* geenidest klasterdatud transkriptsiooniühikutesse, kuna RegulonDB-st 1999.a. üle võetud geneetilise võrgustiku andmed on puudulikud ja neile tehakse hetkel korrektuuri.

EcoCyc-s on kirjeldatud palju membraantransportsüsteeme. Membraantransporterid on vastutavad metaboliitide impordis, metaboolsete lõpp-produktide ekspordis ning osalevad teistes rakuprotsessides (Karp, Riley, Saier *et al*., 2002). Iga transporteri jaoks on EcoCyc-s annotatsioon, mis sisaldab järgmist (Karp *et al*., 2000):

1. funktsionaalne informatsioon (substraadi spetsiifilisus, relatiivne afiinsus)
2. järjestuse sarnasus teiste teadaolevate transporteritega
3. transporteri või tema substraadi füsioloogiline roll metabolismis
4. funktsiooni leidmisel kasutatud eksperimendi lühikirjeldus, näiteks kloneerimine ja ekspressiooni andmed, *knock-out* mutandid, valgu puhastamine jne.
5. lisainformatsioon – domääni struktuur, transkriptsiooni regulatsiooni detailid jne.

EcoCyc andmebaas sisaldab lisaks detailseid kirjeldusi ensüümi poolt katalüüsitud reaktsioonidest, ensüümi poolt aktsepteeritud substraadi kogustest, ensüümi aktiveerivatest või inhibeerivatest kemikaalidest ning subühikute struktuuridest (Karp, 2001). *E.coli*

metaboolse võrgustiku visualiseerimiseks on EcoCyc-s välja töötatud visualiseerimistööriist Metabolic Overview, mis on üks osa EcoCyc-i graafilisest kasutajaliidesest (Karp *et al*., 1999). EcoCyc-i ja MetaCyc-i visualisatsioonid on tehtud sama tööriistaga ja näide sellest on joonisel 3 (vt. lk. 18).

4.2.1 Pathway Tools tarkvara

EcoCyc-s on arendatud Pathway Tools tarkvara päringute, analüüside ja visualiseerimiste teostamiseks. Pathway Tools-i võib nimetada funktsionaalse bioinformaatika keskkonnaks. Seega saab EcoCyc-i kasutada organismi genoomi järgi metaboolse võrgustiku ennustamiseks.

Pathway Tools koosneb neljast peamisest komponendist (Karp, Paley *et al*., 2002):

1. PathoLogic toetab uue PGDB (Pathway/Genome Database) loomist organismi annoteeritud genoomi põhjal. PathoLogic poolt loodud andmebaas sisaldab uuritava organismi geene, valke, biokeemilisi reaktsioone ja ennustatavaid metaboolseid radu.
2. Pathway/Genome Navigator pakub päringu-, visualiseerimis- ja analüüsiteenuseid PGDB-le.
3. Pathway/Genome Editors – PGDB-de interaktiivne uuendamine (uue metaboolse raja loomine, transportvalgu funktsiooni muutmine, transkriptsioonifaktori ja tema seostumissaidi vahelise interaktsiooni määramine jne.).
4. Pathway Tools ontoloogia määrab PGDB skeemi (klassid, atribuudid, suhted bioloogiliste andmete – metaboolsed rajad, ensüümide funktsioonid, DNA regioonid, geeniregulatsiooni mehhanismid - modelleerimiseks).

Antud tööriistaga saab luua uue radade/genoomi andmebaasi (PGDB) mikroorganismi kohta, mille genoom on sekveneeritud (Karp, 2001). Radade/genoomi andmebaasidest ning nendega lähedalt seotud mudelorganismi andmebaasidest (MOD) on täpsemalt juttu järgmises peatükis.

4.2.2 Pathway/Genome Databases ja Model Organism Databases

Tervikuks ühendatud radade/genoomi andmebaasid kirjeldavad geene ja organismi genoomi, nende ennustatavaid radasid, reaktsioone, ensüüme ja metaboliite. Samuti saab radade/genoomi andmebaasi abil kirjeldada organismi geneetilist võrgustikku: promooterid, operonid, transkriptsioonifaktorid ja transkriptsioonifaktorite seostumissaidid. Koos visualiseerimis- ja analüüsitarkvaraga võimaldavad nad paremini mõista organismi

füsioloogiat. Mikroobide puhul kasutatakse PGDB-d avastamaks antimikroobseid ravimeid. Meditsiiniga seotud mikroorganismidest on loodud andmebaaside kogu Pangea Systems-i juurde (http://www.pangeasystems.com), kuhu kuuluvad veel nii EcoCyc kui MetaCyc (vt. peatükk 4.3). Pangea Systems-s on metaboolne võrgustik kirjeldatud nelja bioloogilise objekti tüübiga:

1. rajad, mis moodustavad võrgustiku
2. reaktsioonid, mis moodustavad iga raja
3. metaboolsed ühendid (substraadid, aktivaatorid, inhibiitorid)
4. ensüümid, mis katalüüsivad reaktsioone.

Andmebaas ei sisalda hetkel membraantransporti, geeniregulatsiooni ega signaali ülekande radasid. Genoomid on Pangea Systems-s kirjeldatud kolme bioloogilise objekti tüübiga (Karp *et al*., 1999):

1. mikroorganismi sekveneeritud kromosoomide ja plasmiidide genoomikaardid
2. genoomi kuuluvad geenid
3. geenidele vastavad geeniproduktid.

Seega ühendab PGDB endas geenid, valgud, organismi geneetilise ning metaboolse võrgustiku.

Mudelorganismi andmebaasid (MOD) omavad postgenoomsel ajastul tähtsaid funktsioone. Radade/genoomi andmebaasid kuuluvad samuti MOD alla. MOD on (Karp, Paley *et al*., 2002):

1. elektrooniline allikas organismi genoomi järjestusele
2. ühendab mitmed erinevad teadlaste poolt tehtud geenide funktsioonide ennustused
3. allikas organismi funktsionaalgenoomika õppimiseks
4. alus süsteemibioloogia uurimiseks.

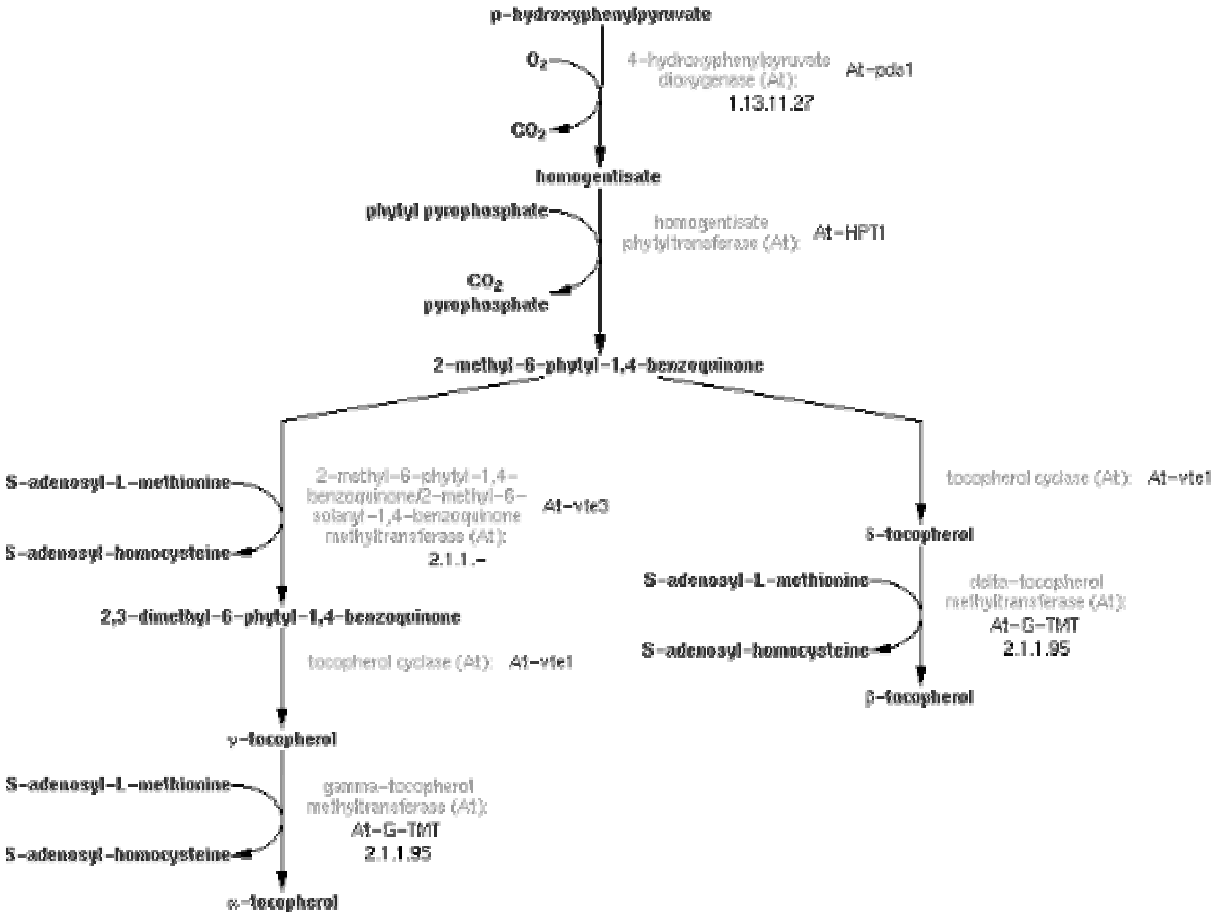
Kuigi praegu on trendiks, et iga MOD arendab oma unikaalse tarkvara ja andmebaasi keskkonna, ei ole see praktiline, sest (Karp, Paley *et al*., 2002):

1. igale uurimisrühmale oma MOD keskkonna loomine on tunduvalt kulukam võrreldes olemasoleva tarkvara kasutamisega
2. MOD keskkonnad sisaldavad niivõrd keerukaid algoritme, et mõned uurimisrühmad ei suuda neid järgida
3. võrdlevad analüüsid üle mitme MOD-i muutuvad raskeks.

MOD on organismi molekulaarsete osade ja nendevaheliste interaktsioonide kogum ning seetõttu oluline süsteemibioloogia teooriate säilitamiseks, testimiseks ja edasi arendamiseks (Karp, Paley *et al*., 2002).

4.3 MetaCyc

MetaCyc (http://metacyc.org) on metaboolsete radade andmebaas, kus rajad on eksperimentaalselt tuvastatud ning pärinevad valdavalt mikroorganismidelt ja taimedelt. Enim uuritud organism on *E. coli* (173 rada), inimesel on kirjeldatud 31 metaboolset rada. Kokku on MetaCyc-s teaduslikult kirjeldatud enam kui 240 organismi metaboolsed rajad. Iga rada on märgistatud liigi nimega, kus ta esineb.



Joonis 3. MetaCyc-i näide: E-vitamiini biosünteesirada

MetaCyc-i peaeesmärk on metabolismi universumi kaardistamine. Rakendused: mikroobi ja taime radade ulatuslik allikas, genoomi järjestuse põhjal organismi metaboolsete radade ennustamine, genoomide radade analüüs, metaboolne tehnoloogia, biokeemiliste võrgustike võrdlemine, süsteemibioloogia. MetaCyc kasutab sama andmebaasiskeemi, mis EcoCyc, kuid ei sisalda genoomi kaarte (järjestusi). MetaCyc sisaldab viiteid mitmetesse andmebaasidesse: SWISS-PROT, PIR (Protein Information Resource) (Wu *et al.*, 2003), PDB (Westbrook *et al.* 2002), TAIR (Rhee *et al.*, 2003), SGD (Dwight *et al.*, 2004).

MetaCyc klassid: metaboolsed rajad, reaktsioonid, ensüümid, ühendid, viited. Nimetatud andmebaas sisaldab ulatuslikku informatsiooni ensüümide kohta: aktivaatorid,

inhibiitorid, kofaktorid, prosteetilised rühmad[8], alternatiivsed substraadid, selgitavad kommentaarid, viited. Samuti võimaldab ta otsida ensüüme ja radasid ensüümi katalüütilise funktsiooni järgi. Lisaks sisaldab MetaCyc erinevaid taksonoomiaid: EC süsteem, radade klasside ja ühendite klasside hierarhia (Karp, Riley, Paley *et al.*, 2002; Krieger *et al.*, 2004).

4.3.1 MetaCyc *versus* KEGG

1. Erinevalt KEGG-st sisaldab MetaCyc ulatuslikke kommentaare radade ja ensüümide kohta.
2. MetaCyc viitab kirjanduslikele allikatele, kust radade ja ensüümide andmed pärinevad, KEGG-s vastavad andmed puuduvad.
3. MetaCyc rajad on väiksemad, sest KEGG kombineerib kokku erinevate liikide samad rajad.
4. MetaCyc rajad on märgistatud infoga, millisel liigil antud rada on eksperimentaalselt kindlaks tehtud, KEGG-s selline info puudub. Samas võimaldab KEGG kasutajal näha, millised ensümaatilised sammud ennustatavalt toimuvad antud rajal mitmetes organismides.
5. Erinevalt KEGG-st sisaldab MetaCyc andmeid spetsiifiliste ensüümide omaduste kohta eri liikides, näiteks alamühiku ülesehitus, substraadi spetsiifilisus, kofaktori nõuded, aktivaatorid, inhibiitorid. (Karp, Riley, Paley et al., 2002).

4.4 aMAZE

aMAZE (http://www.amaze.ulb.ac.be) („a maze“ – interaktsioonide labürint molekulaarsete sündmuste vahel rakus) relatsiooniline andmebaas sisaldab informatsiooni geenieksprессiooni, katalüüsitavate keemiliste reaktsioonide, regulatoorsete interaktsioonide, valgu perekondade ning metaboolsete ja signaali ülekande radade kohta. aMAZE on arendatud eesmärgiga luua mahukas infoallikas rakus toimuvatest interaktsioonidest ja protsessidest. Kuna põhiinfo bioloogilise funktsiooni kohta on järjestuste andmebaasides (näiteks SWISS-PROT, GenBank (Benson *et al.*, 2004)) ning sealne tekstiline kirjeldus ei ole kohandatav arvutusmeetoditeks, kujutatakse aMAZE-s bioloogilist funktsiooni molekulaarsete sündmuste ja protsesside kaudu.

[8] prosteetiline rühm – valkaine molekuli mittevalguline osis

aMAZE keskkond koosneb kahest suuremast komponendist. aMAZE LightBench on nimetatud andmebaasi veebiliides. Sealt saab pärida keemiliste reaktsioonide, metaboolsete radade olevate geenide ja ensüümide, valk-valk interaktsioonide ning valgu modifikatsioonide kohta signaali ülekande radades. Teine oluline komponent on aMAZE WorkBench (Java-põhine rakendus), mis võimaldab andmete laadimist ja pakkimist, modifikatsioone ja annotatsioone, visualiseerimist ja analüüsi (Lemer et al., 2004).

aMAZE andmemudelis on kaks peaklassi: *Biochemical Entities* ning *Biochemical Interactions*. *Biochemical Entities* kirjeldab struktuurseid üksusi (valk, geen, keemiline ühend jne.) atribuutidega, mis kuuluvad struktuursete omaduste juurde (näiteks geeni asukoht kromosoomis). *Biochemical Interactions* kirjeldab erinevat tüüpi molekulaarseid sündmusi (reaktsioon, ekspressioon, transkriptsiooni regulatsioon jne.). Kolmas tähtis klass on *Process*, mis sisaldab nii individuaalseid interaktsioone kui terveid protsesse (Lemer *et al.*, 2004).

4.5 Reactome

Reactome (endise nimega Genome Knowledgebase - GKB) on inimese bioloogiliste protsesside andmebaas, mis katab rajad alates metabolismi põhiprotsessidest (näiteks glükolüüs) kuni kõrgema taseme protsessideni välja (näiteks hormonaalsed signaalid). Reactome hõlmab endas järgnevaid bioloogilisi protsesse: rakutsükkel, DNA reparatsioon ja replikatsioon, geenikspressioon, mRNA protsessimine ja translatsioon ning suhkrute, etanooli, aminohapete, nukleotiidide ja lipiidide metabolism. Reactome omab viiteid GO (http://www.geneontology.org), Ensembl (Birney et al., 2004), ChEBI (http://www.ebi.ac.uk/chebi/) ja Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/) andmebaasidesse. Reactome-i tööriist Skypainter võimaldab reaktsioonikaardi värvimist vastavalt kasutaja poolt märgitud geenide või valkude identifikaatoritele. Teine tööriist Pathfinder teeb kindlaks rajad, mis ühendavad kasutaja poolt märgitud sisendeid ja väljundeid. Kui on märgitud mitu väljundmolekuli, arvutab Pathfinder lühima tee antud molekulide vahel. Kuigi Reactome on suunatud inimese bioloogilistele radadele, sisaldab ta mitmeid individuaalseid biokeemilisi reaktsioone teistelt organismidelt nagu hiir, rott, sebrakala, fugu ja kana (Joshi-Tope *et al.*, 2005).

4.5.1 Reactome-i andmemudel

Reactome-i autorid on koostanud mahuka hierarhilise andmemudeli (http://www.reactome.org/cgi-bin/classbrowser – allpool toodud peaklasside ja teiste kommentaaride paremaks mõistmiseks). Mudeli põhiühikuks on reaktsioon. Reactome eeldab, et kui sisendid ja ensüümid on olemas, siis reaktsioon toimub. Reaktsioon on sündmus, mis konverteerib sisendeid väljundeiks, kus sisenditeks ja väljunditeks on molekulid, valgud, lipiidid, nukleotiidid või nende kompleksid. Reaktsioon sisaldab lisaks informatsiooni liigi, subtsellulaarse asukoha ja eksperimentaalsete tingimuste kohta. Reaktsioonid grupeeritakse radadesse ning need omakorda võivad moodustada võrgustiku.

Vaatamata välisele keerukusele on Reactome-s üllatavalt väike kogus tuumkontseptsioone – neli peaklassi: *Event, Physical Entity, ReferenceEntity ja Catalyst Activity* (Joshi-Tope *et al.*, 2003).

Event. *Event* klass kirjeldab rakulisi protsesse. *Event* on tabel, kus asuvad Reactome tekstilised kirjeldused, mis on olulised sündmuste detailide täpsustamiseks. *Event* omab kahte alamklassi - *Pathway* ja *Reaction*. *Pathway* kirjeldab funktsionaalselt seotud reaktsioonide gruppe. *Reaction* kirjeldab individuaalseid biokeemilisi reaktsioone, aga ka molekulaarsete komplekside moodustumist ja lagunemist ning molekulide transporti ühest raku asukohast teise. *Reaction* tabelis eristatakse selliseid reaktsioone, millel on samad sisendid ja väljundid, kuid erinev katalüsaator (http://www.reactome.org).

PhysicalEntity. Need on kõik kompleksid, molekulid, ioonid ja partiklid Reactome-s. Antud klass hõlmab ka molekulide struktuure ja rakusiseseid asukohti. Eristatakse sama järjestusega valke, millest üks on fosforüleeritud, aga teine on modifitseerimata. Samuti on eristatud keemiliselt identsed molekulid, mis asuvad erinevates raku subtsellulaarsetes asukohtades (näiteks ATP tsütosoolis ja ATP mitokondri maatriksis). *PhysicalEntities* on omakorda jaotatud mitmesse alamklassi (http://www.reactome.org).

CatalystActivity. *CatalystActivity* klass määrab kindlaks keemilise ühendi, mis võtab katalüüsitavast reaktsioonist osa, molekulaarse funktsiooni (või tõenäoliselt mitu funktsiooni). Samuti täpsustatakse, kas katalüüsist võtab osa kogu kompleks või ainult kompleksi kindel alamühik. Vajadus nimetatud klassi järele seisneb selles, et mõned kompleksid või valgud omavad rohkem kui üht aktiivsust. Seega on võimalik lühidalt kirjeldada mitmeid võimalikke reaktsioone (http://www.reactome.org).

ReferenceEntity. Need on alamüksusteta molekulid, modifikatsioonideta valgud, RNA ja DNA molekulid ning keemilised ühendid. *ReferenceEntity* sisaldab antud molekulide nimesid, struktuurseid omadusi ja viiteid teistesse andmebaasidesse. Sarnaselt eelpool kirjeldatud klassidele omab ka *ReferenceEntity* mitmeid alamklasse (http://www.reactome.org).

4.5.2 Konkreetne versus Üldine (Concrete vs Generic)

Paljudel Reactome klassidel on kaks kuju, konkreetne ja üldine (*Concrete* ja *Generic*). *Concrete* klassiga on seotud spetsiifilised andmed, näiteks kindel valk konkreetses liigis viib läbi spetsiifilist reaktsiooni. *Concrete* klassid kuuluvad alati *Generic* klasside hulka. *Generic* klassid on mõeldud rohkem üldiste sündmuste jaoks nagu fosforüleerimine, mida mitmed valgud erinevates organismides läbi viivad. Reactome kasutab *Generic* klasse kirjeldamaks sama mõistet erinevatel liikidel, näiteks ribosoom (*Generic Complex*) või tsitraaditsükkel (*Generic Pathway*). *Generic* klassid omavad peaaegu alati seoseid GO „molekulaarse funktsiooni", „bioloogilise protsessi" ja „rakulise komponendi" hierarhiates ning Reactome otseselt kasutab GO termineid igal võimalikul hetkel (Joshi-Tope *et al.*, 2003).

5. Metaboolsete radade visualiseerimine

Metaboolsete radade puhul annab korrektne visualisatsioon peaaegu sama palju informatsiooni kui algebralised või eksperimentaalsed andmed tabeli kujul. Radade graafika ja topoloogia võib sisaldada mittetekstilises vormis piisavalt informatsiooni radade funktsioneerimise mõistmiseks. Graafiline visualiseerimine peab olema informatsioonitihe, ülekoormamata, äratuntav ja ühtlane. Tänapäeval on välja töötatud mitmed süsteemid radade kolmedimensionaalseks (3D) visualiseerimiseks, mis on heaks lisamaterjaliks tavalistele 2D skeemidele. 3D visualiseerimiseelised võib kokku võtta järgmiselt (Rojdestvenski, 2003):

1. Infotihedus – visuaalsete atribuutide kasutamine suurendab graafiku infosisaldust
2. Ülevaatlik ja detailne vaade – kasutajale on antud liikumisvabadus virtuaalses maailmas. Kui huviobjektiks on kindel piirkond, siis seda on võimalik näha detailselt. 3D maailmas saab erinevaid radade osi fokusseerida, säilitades kogu raja struktuuri.
3. Standard GUI (Graphical User Interface) – kasutades standardseid programmeerimiskeeli ja GUI-d, on kasutajal lihtsam kohaneda süsteemi tööga.

Radade 3D ruumis visualiseerimiseks kasutatakse VRML (Virtual Reality Modeling Language) süsteemi, mis baseerub MNV (Metabolic Network Visualizer) keelel. Süsteemil on

neli komponenti: MNV keele standard ja analüüsija, MNV → VRML translaator, interaktiivne radade ehitaja ning konverterid olemasolevatele radade andmebaasidele (MetaCyc jt.). Kõik süsteemi komponendid on kirjutatud Perlis, HTMLis ja Javascriptis. Keeled määrati sobivuse järgi, sest Perl sobib grammatilise teksti analüüsiks. Võrgustiku MNV-kirjelduse visualiseerimine on tehtud automaatselt ning kogu info visualiseerimise ja eriti graafiku kujutamise jaoks on saadaval MNV kirjelduses. Tähelepanuväärne sarnasus on MNV-l SBML-ga (System Biology Markup Language). MNV on SBMLi alamüksus ning on laiendatav, kuid on loodud eelkõige rahuldamaks visualiseerimisvajadusi. SBML on rohkem üldisem ja komplekssem keel ning sobiv modelleerimiseks. Hetkel tegeletakse uue staadiumi ehk kogu radade kollektsiooni visualiseerimisega 3D-s, näidates interaktsioone eri radade vahel (Rojdestvenski, 2003).

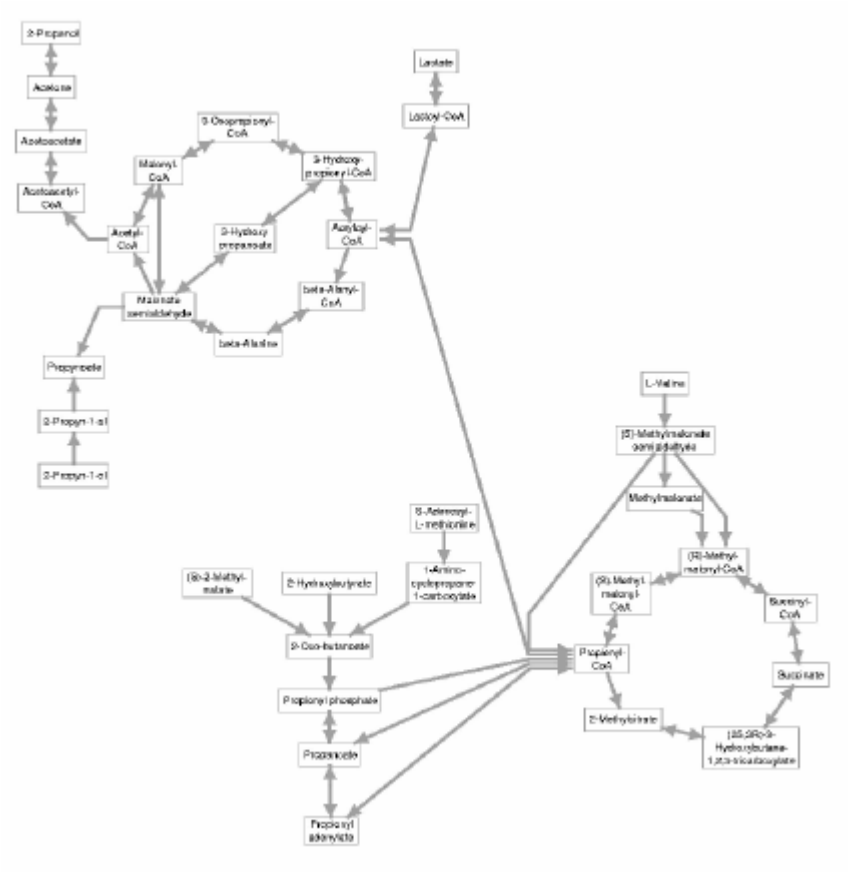
5.1 Staatiline ja dünaamiline visualiseerimine

Eelpool kirjeldatud andmebaasidest visualiseerib KEGG radu staatiliselt - need on käsitsi joonestatud. Staatilise visualiseerimise puudused:

1. andmete uuendamisel tuleb vastavad kujutised käsitsi muuta
2. näidatavate detailide hulka ei ole võimalik määrata ega radade osi peita
3. kui on vaja visualiseerida kasutaja poolt määratud või uudseid radu, siis ei ole staatiline visualiseerimine üldse rakendatav.

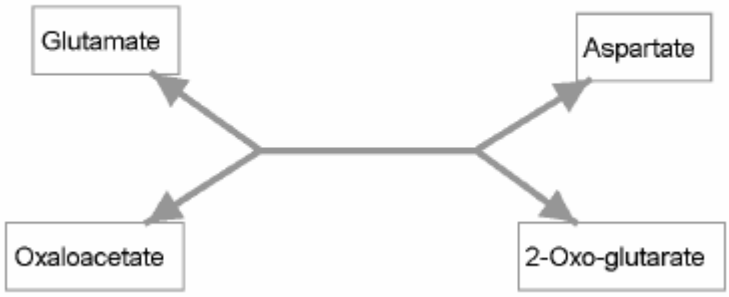
Sellele vastukaaluks pakub dünaamiline visualiseerimine suurt paindlikkust, mis on vajalik komplekspäringute jaoks ja uute radade konstrueerimiseks. Saades sisendiks graafi matemaatilise kirjelduse, peab dünaamilise visualiseerimise algoritm graafi elemendid geomeetriliselt paika panema kindlate reeglite järgi. EcoCyc-i loojad Karp ja Baley märkisid, et tuleb kasutada eri algoritme tsükliliste, lineaarsete ja puu-struktuuriga radade jaoks (rakendatud EcoCyc-is) (Karp & Paley, 1994). Sarnane näide on joonisel 4, kus on ühendatud tsüklilised ja lineaarsed rajad.

Metaboolseid radu modelleeritakse tavaliselt suunatud graafidena. Rada koosneb omavahel ühendatud biokeemilistest reaktsioonidest. Substraate ja produkte kujutatakse graafi tippudena ning reaktsioone graafi servadena. Kõrvalsubstraadid joonestatakse serva lähedusse ning ühendatakse servaga kaare abil. Kui reaktsioon on ensüümi poolt katalüüsitud, siis märgistatakse serv ensüümi nimega.



Joonis 4. Skeem, kus on ühendatud tsüklilised ja lineaarsed rajad

Kui tavalised graafiku kujutamise algoritmid joonistavad graafiku kindlate kriteeriumide järgi nagu planaarsus, minimaalne servade ristumine, minimaalne joonestamisala ja maksimaalne sümmeetria, siis metaboolsete radade puhul on raske selliseid reegleid luua. Mõned keemilised reaktsioonid omavad mitut substraati või produkti ning sel juhul võetakse kasutusele hüperservad ehk servad, millel on mitu tippu (joonis 5) (Becker & Rojas, 2001).



Joonis 5. Näide hüperservadest

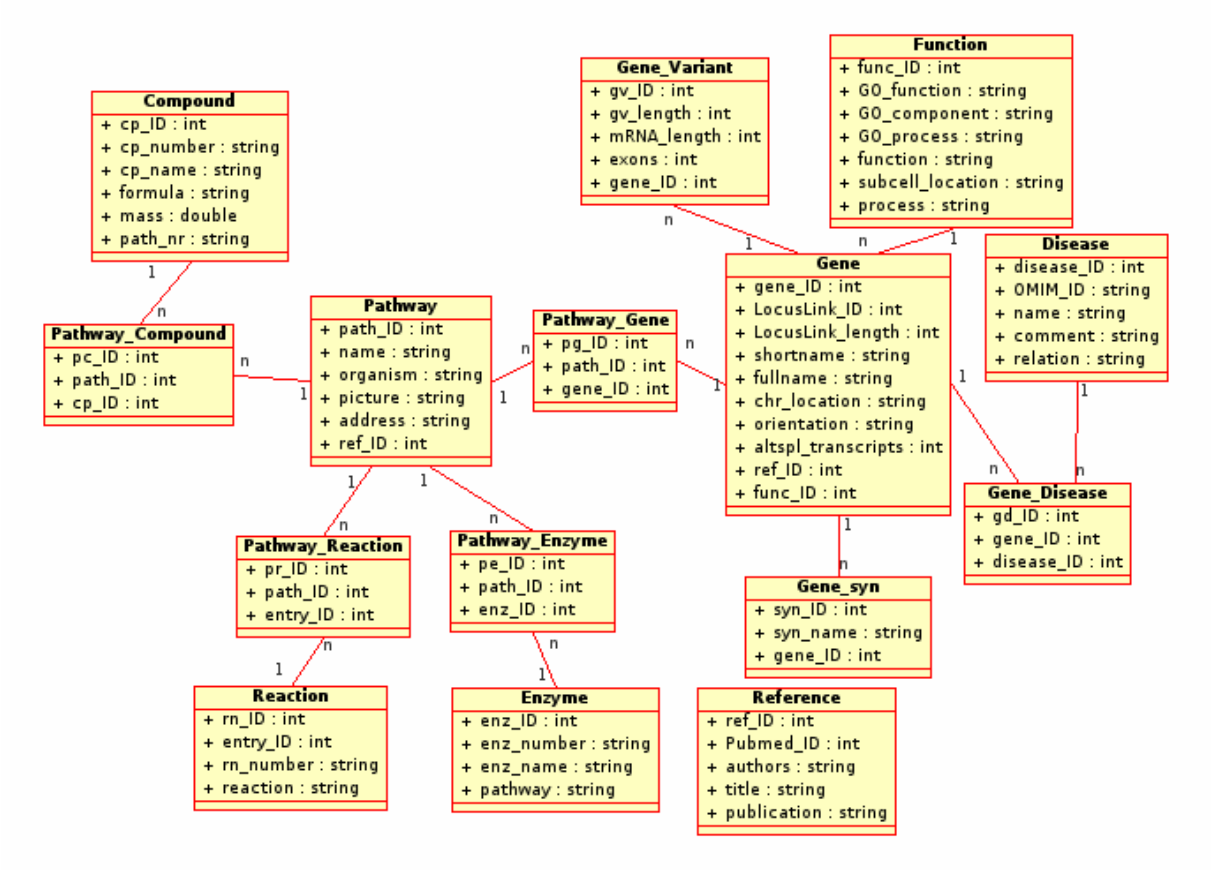
II Praktiline töö

6. Töö eesmärgid

Käesoleva töö eesmärgiks oli uurida metaboolsete radade andmebaaside ülesehitust, tööriistu, nende eeliseid ja puudusi ning luua omapoolse vaatega sarnane andmebaas. Iga andmebaasi loomine algab olemasolevatele andmetele tuginedes sobiva mudeli väljatöötamisest. Antud töö puhul oli vaja adekvaatselt kirjeldada ja omavahel siduda erinevad bioloogilised olemid: bioloogilised rajad, keemilised reaktsioonid, reaktsioone katalüüsivad ensüümid, reaktsioonides osalevad keemilised ühendid ning geenid oma karakteristikutega (nimetus, funktsioon, pikkus jne.). Üks esimesi mahukamaid etappe oligi mudeli lõpliku kuju väljatöötamine. Peale mudeli väljatöötamist olid järgmisteks olulisteks etappideks andmebaasi programmeerimine, andmete töötlemine ühtsele kujule ning sisestamine andmebaasi. Töö viimase osana on valminud veebiliides, kust saab otsida infot geenide, radade, ensüümide ja keemiliste ühendite kohta vastavalt märksõna järgi.

7. Andmebaasi struktuur

Peale tutvumist teiste sarnaste bioloogiliste radade andmebaasidega kirjanduse ja veebi kaudu, koostasime 15-st tabelist koosneva andmemudeli, mille peamised tabelid on *Gene, Pathway, Reaction, Enzyme, Compound* ja *Disease*. Andmebaasi mudel on koostatud Umbrello programmiga ja on kujutatud joonisel 6. Mudeli esmasel loomisel on aluseks võetud bioloogiliste radade andmed ehk kesksele kohale asetusid tabelid *Gene* ja *Pathway* oma atribuutidega. *Pathway* tabel seob omavahel reaktsioonid, keemilised ühendid, ensüümid ja geenid ning *Gene* omakorda on ühendatud *Disease* tabeliga. Geenide ekson/intron struktuuri ning mRNA pikkusi kirjeldavad andmed on paigutatud *Gene* tabelist eraldi *Gene_Variant* tabelisse, kuhu saab vajadusel lisada alternatiivselt splaissitud geene, sest needki on seotud mitmete haigustega (rinnavähk, meeste viljatus jt.). *Disease* tabelis on hetkel hüpertensiooni erinevate vormide nimetused (mõnede puhul ka sünonüümid) ning kommentaarid haiguse kohta. Lisaks on andmemudelis viis peatabelite sidumiseks mõeldud vahetabelit. Andmemudeli tabelite ja nende atribuutide täpsemad kirjeldused on esitatud järgmises peatükis.



Joonis 6. Radade andmebaasi mudel

8. Tabelite ja nende atribuutide kirjeldus

Tabelite kirjeldamisel jätan välja vahetabelid ning atribuutidest ei käsitle erinevaid identifikaatoreid, mis on igale tabelile unikaalsed.

Pathway

name - raja nimi, näiteks Glycolysis, Galactose metabolism

organism – liik, kus vastav rada on kindlaks tehtud (hetkel ainult inimese rajad)

picture – raja tunnus (kasutatakse sidumiseks)

address – viide KEGG andmebaasi skeemile antud rajast

Reaction

entry_ID – reaktsiooni tunnus

rn_number – raja tunnus (kasutatakse sidumiseks)

reaction – reaktsiooni kirjeldus

Compound

cp_number – keemilise ühendi tunnus

cp_name – keemilise ühendi nimi

formula – keemilise ühendi valem

mass – keemilise ühendi molekulmass (kDa[9])

path_nr – raja tunnus (kasutatakse sidumiseks)

Enzyme

enz_number – EC (Enzyme Commission) number

enz_name – ensüümi nimi

pathway – raja tunnus (kasutatakse sidumiseks)

Gene

LocusLink_ID – geeni identifikaator LocusLink andmebaasis

LocusLink_length – geeni pikkus (bp) LocusLink andmebaasi järgi

shortname - geeni lühend, näiteks ACE, CLCNKA jne.

fullname - geeni täisnimetus, näiteks Angiotensin-II type 1 receptor

chr_location - geeni kromosomaalne lokalisatsioon, näiteks 11p32.1

orientation - geeni orientatsioon (+/- ahel)

altspl_transcripts - alternatiivse splaissingu transkriptide arv

Gene_Variant

gv_length - geeni variandi pikkus

mRNA_length - geeni poolt kodeeritud mRNA pikkus

exons - geeni ekson/intron struktuur (mitu eksonit)

Function

GO_function – GO Molecular Function, näiteks GO:0016209

GO_component – GO Cellular Component, näiteks GO:0005623

GO_process – GO Biological Process, näiteks GO:0030534

function - geeni funktsioon

[9] kDa – kilodalton, molekulmassi ühik (1kDa = 1000 Da)

subcell_location - geeni subtsellulaarne lokalisatsioon, näiteks nuclear (tuum)
process - bioloogiline protsess, millest geen osa võtab

Gene_syn

Tabel geeni sünonüümide jaoks. Lisaks sünonüümi ja geeni ID-le on tabeli kolmandaks atribuudiks syn_name ehk sünonüümide nimed.

Disease

OMIM_ID – haiguse id OMIM (Online Mendelian Inheritance in Man) andmebaasis
name – haiguse nimi, näiteks Hypertension, essential, salt-sensitive
comment - kommentaar haigusele
relation – geenide omavahelised suhted haiguse tekkel

Reference

Pubmed_ID – kirjandusliku viite identifikaator PubMed andmebaasis
authors - artikli autorid
title - artikli pealkiri
publication - ajakirja nimi, number, leheküljed, aasta

9. Andmete sisestamine ja pärimine

Andmete töötlemiseks ja andmebaasi sisestamiseks on loodud Perlis kirjutatud programmid vastavalt geenide, keemiliste ühendite, ensüümide ja radade andmete jaoks. Kõik programmid sisestavad andmebaasi andmeid tekstifailidest.

Näidispäringud on toodud erinevatest valdkondadest, mis on kõik antud andmebaasis esindatud.

1. Päringu näide radade valdkonnast:

```
SELECT path_ID, name, organism, address
FROM Pathway
WHERE name LIKE '%signal%';
```

Päringu tulemuseks on kõigi radade, kus esineb sõna signal, ID-d, nimed, organism, kus esineb (inimene) ja viited KEGG andmebaasi skeemidele.

Name	Organism	Address
MAPK signaling pathway	H.sapiens	http://www.genome.ad.jp/dbget-bin/show_pathway?hsa04010
Calcium signaling pathway	H.sapiens	http://www.genome.jp/dbget-bin/show_pathway?hsa04020
Phosphatidylinositol signaling system	H.sapiens	http://www.genome.ad.jp/dbget-bin/show_pathway?hsa04070
Wnt signaling pathway	H.sapiens	http://www.genome.ad.jp/dbget-bin/show_pathway?hsa04310
Notch signaling pathway	H.sapiens	http://www.genome.jp/dbget-bin/show_pathway?hsa04330
TGF-beta signaling pathway	H.sapiens	http://www.genome.ad.jp/dbget-bin/show_pathway?hsa04350
Toll-like receptor signaling pathway	H.sapiens	http://www.genome.ad.jp/dbget-bin/show_pathway?hsa04620
Jak-STAT signaling pathway	H.sapiens	http://www.genome.ad.jp/dbget-bin/show_pathway?hsa04630

Nagu eelpool mainitud, viitab *address* antud raja skeemile KEGG andmebaasis. Kui üldiselt on andmebaasides eelistatumad lihtsamad võrgustikud, siis KEGG-s on rõhutud infotihedusele ja seetõttu on sealsetel skeemidel raskem orienteeruda. Samas on KEGG-i skeemid korrapärased ja sümmeetrilised ning heaks aluseks spetsiifiliste visualisatsioonide loomisel.

2. Geenide osalus erinevate haigustega ehk päritakse haiguse märksõna järgi ja kui see märksõna esineb haiguse või haiguse kommentaari tulbas, saame päringu tulemusena info vastavate geenide kohta.

Päring:

```
SELECT Gene.shortname, Gene.chr_location, Gene.LocusLink_length, Disease.name
FROM Gene, Disease, Gene_Disease
WHERE Gene.gene_ID = Gene_Disease.gene_ID AND Gene_Disease.disease_ID =
Disease.disease_ID AND (Disease.name LIKE '%deficiency%' OR Disease.comment LIKE
'%deficiency%');
```

Antud päringu tulemusena saame teada kõikide geenide lühendid, kromosomaalsed asukohad, LocusLink pikkused ja haiguste nimed, kui haiguse nimes või kommentaaris esineb sõna *deficiency*.

shortname	chr_location	LocusLink_length	name
CYP11B1	8q21	7464	Adrenal hyperplasia, congenital, due to 11-beta-hydroxylasedeficiency (3); Aldosteronism, glucocorticoid-remediable
CYP11B2	8q21-q22	7284	Hypoaldosteronism, congenital, due to CMO II deficiency (3);Hypoaldosteronism, congenital, due to CMO I deficiency, 203400 (3); {Low renin hypertension, susceptibility to} (3); Aldosterone to renin ratioraised (3)
CYP17A1	10q24.3	6886	Adrenal hyperplasia, congenital, due to 17-alpha-hydroxylasedeficiency (3)
HSD11B1	1q32-q41	30077	Cortisone reductase deficiency, 604931 (3)
APOA1	11q23-q24	1869	1. ApoA-I and apoC-III deficiency, combined; 2. Hypertriglyceridemia
APOA2	1q21-q23	1335	Apolipoprotein A-II deficiency; hypercholesterolemia, familial
APOA4	11q23	2602	Apolipoprotein A-II deficiency; hypercholesterolemia, familial
CETP	16q21	21994	CETP deficiency. Probably involved in the development od atherosclerosis.
GH1	17q24.2	1635	Isolated growth hormone deficiency, IIIig type with absent GH.
SELP	1q22-q25	41	Platelet alpha/delta storage pool deficiency
TBXAS1	7q34-q35	191	Thromboxane synthase deficiency.

3. Haiguste täpsemad kirjeldused (hetkel ainult hüpertensiooni eri vormide kohta). Haiguste kirjelduste jaoks on loodud *Disease* tabelisse *comment* atribuut. Lisaks on *Disease* tabelis viidatud OMIM andmebaasile, et vajadusel sealt edasi otsida.

Päring:

```
SELECT name, comment
FROM Disease
WHERE name LIKE '%Bartter%';
```

Antud päringu tulemuseks saame kõigi haiguste nimed ja kommentaarid, kus esineb sõna Bartter.

name	comment
Bartter syndrome, infantile, with sensorineural deafness	
Bartter syndrome, antenatal, 601678 (3); Bartter syndrome, 241200 (3)	Defects in CLCNKB are a cause of type 3 Bratter syndrome, an autosomal recessive form of often severe intravascular volume depletion due to renal salt-wasting associated with low blood pressure, hypokalemic alkalosis, hypercalciuria, normal serum magnesium levels.
Bartter syndrome characterized by an hypokalemic, hypocholemic metabolic alkalosi with hyperkaliury, hyperexcretion of prostaglandin E, hyperreninemia hyperaldosteronism with normal blood pressure, intensitivity to AGT2, and hyperplasia of juxtaglomerular	Defects in KCNJ1 are the cause of antenatal Bartter syndrome (ABS). ABS is a phenotypically distinct variant of the Bartter syndromes characterized by polyhydraminos with preterm delivery, severe salt wating, hyposthenuria, and hypercalcuria. ABS is a life-threatening disorder in which both renal tubular hypokalemic alkalosis as well as profound systemic symptoms are manifest. ABS is also termed hyperprostangladin E syndrome.
Bartter syndrome, hypokalemic, hypochloremic metabolic alkalosis with hyperkaliury, hyperexcretion of prostaglandin E, hyperreninemia hyperaldosteronism with normal blood pressure, intensitivity to AGT2, and hyperplasia of juxtaglomerular apparatus, autos	Defects in SLC12A1 are a cause of Bartter syndrome (BS). BS is a life-threatening condition beginning in utero, with marked fetal polyuria that leads to polyhydramnios and premature delivery. Another hallmark of this disease is a marked hypercalciuria and, as a secondary consequence, the devalopment of nephrocalcinosis and osteopenia.

4. Andmebaasi saab kasutada ka EC numbri järgi ensüümi nime ja vastupidiselt ensüümi nime järgi EC numbri otsimiseks. Järgnev päring annab tulemuseks kõikide ensüümide, mille EC number sisaldab kombinatsiooni 1.7.3, täielikud EC numbrid ja nimed.

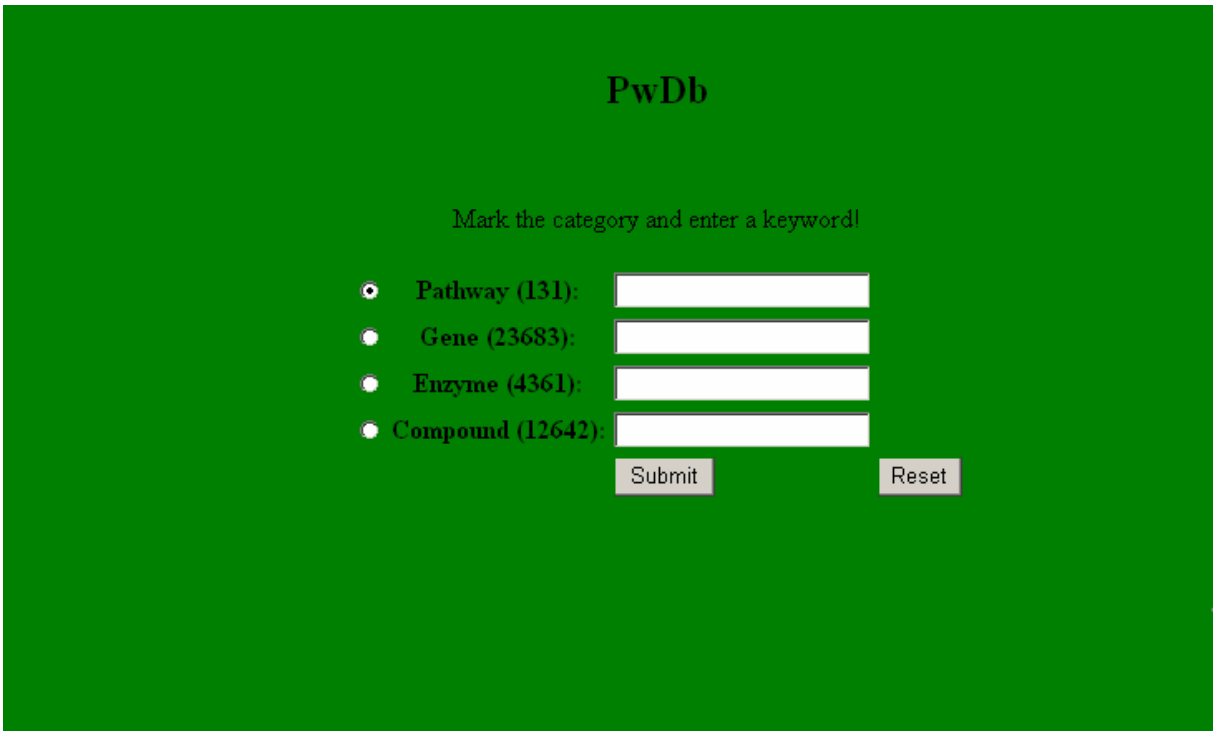
Päring:

```
SELECT enz_number, enz_name
FROM Enzyme
WHERE enz_number LIKE '%1.7.3%';
```

EC Number	Enzyme
EC 1.7.3.1	nitroethane oxidase
EC 1.7.3.2	acetylindoxyl oxidase
EC 1.7.3.3	urate oxidase uricacidoxidase
EC 1.7.3.4	hydroxylamine oxidase HAO
EC 1.7.3.5	3-aci-nitropropanoate oxidase propionate-3-nitronateoxidase
EC 3.1.7.3	monoterpenyl-diphosphatase bornylpyrophosphatehydrolase

10. Andmebaasi veebiliides

Andmebaasi laiemaks kasutamiseks on loodud veebiliides, mis asub aadressil http://kotkas.ebc.ee/u/raudsepp. Veebiliides on kirjutatud HTML ja EmbSQL keeltes. Embsql on perli moodul, mis on loodud SQL-i päringute tegemiseks ning päringutulemuste näitamiseks veebilehtedel. Päringute tegemiseks tuleb vastav kategooria (Pathway, Gene, Enzyme, Disease) märgistada ning sisestada märksõna. Iga kategooria järel on sulgudes kirjas, kui palju kirjeid antud kategooria sisaldab.



Joonis 7. Andmebaasi veebiliides.

Päringu tulemused kuvatakse tabeli kujul. Kui märksõnale vastavat tulemust ei eksisteeri, näeb kasutaja ainult tabeli päist vastavate veergude nimedega. Radade päringu korral sisaldab tulemustabel nime, organismi ja KEGG andmebaasi viidete veerge. Pärides geeni nime järgi, saab tulemuseks geeni lühendi, täisnime, kromosomaalse asukoha ja LocusLink identifikaatori. Ensüümi päringu tulemuseks on ensüümi täielikud nimed ja EC numbrid. Keemilise ühendi päring annab tulemuseks ühendi täieliku nime, keemilise valemi ning molekulmassi.

Näidisena on toodud päring, kus keemilise ühendi märksõnaks on *sulo*. Päringu tulemus on kuvatud joonisel 8, kus on välja toodud kõik antud märksõna sisaldavad keemilised ühendid koos valemi ja massiga.

Name	Formula	Mass
Sulochrin	C17H16O7	332.0895
5-Dehydro-4-deoxy-D-glucuronate; 4-Deoxy-L-threo-5-hexosuloseuronate	C6H8O6	176.0321
Tamsulosin	C20H28N2O5S	408.1719
Tamsulosin hydrochloride; Flomax(TN)	C20H28N2O5S.HCl	444.1486
Cefsulodin	C22H20N4O8S2	532.0723
Cefsulodin sodium; Takesulin(TN)	C22H19N4O8S2.Na	554.0542

Joonis 8. Alamstringi *sulo* sisaldavad keemillised ühendid.

11. Andmebaasi statistika

Hetkeseisuga on andmebaasis info 131 teadlaste poolt uuritud inimese metaboolsete ja regulatoorsete radade ning üksikute haiguslike radade (Parkinsoni tõbi, Alzheimeri tõbi jt.) kohta. Need jaotuvad omakorda järgnevalt: 106 metaboolset, 11 signaali ülekande, 7 rakuliste protsesside ja 7 haiguslikku rada. Andmebaasis on lisaks kirjeldatud 4361 ensüümi, 12642 keemilist ühendit ja 23683 geeni. Kogu info pärineb KEGG andmebaasist.

KOKKUVÕTE

Töö teoreetilises osas on esmalt välja toodud bioloogiliste radade andmebaaside eesmärgid ja väljakuktsed. Seejärel on pikemalt kirjeldatud erinevaid valk-valk interaktsioonide ja metaboolsete radade andmebaase ning GO konsortsiumi, mis tegeleb ühtse terminoloogia loomisega andmebaaside vahel.

Kuna iga rakuline protsess on reguleeritud valk-valk interaktsioonide poolt, siis nende tuvastamiseks on väljatöötatud mitmeid eksperimentaalseid meetodeid. Need on omakorda aluseks valk-valk interaktsioonide andmebaaside loomisel ja levikul. Bioloogiliste radade andmebaasid on vajalikud genoomi analüüsiks, sest nad kirjeldavad geene ja organismi genoomi, ennustatavaid radasid, reaktsioone, ensüüme ja metaboliite. Koos visualiseerimis- ja analüüsitarkvaraga võimaldavad nad paremini mõista organismi füsioloogiat. Teadlased püüavad ennustada valkude interaktsioonivõrgustikke, mis on vastutavad erinevate protsesside eest rakus. Edu selles valdkonnas nõuab arusaamist arvutiteadusest, genoomikast ja mõõtmistehnoloogiatest, lisaks on suureks väljakutseks geeniregulatsiooni loogika ja biokeemiliste võrgustike identifitseerimine. Peaeesmärgiks on andmebaasi autoritel kogu raku biokeemilise võrgu kujutamine arvutis.

Antud töö sissejuhatuses sai ülesandeks seatud luua andmebaas inimese bioloogilistest radadest, mis hõlmaks ka keemilisi reaktsioone, geene, ensüüme, keemilisi ühendeid ja haigusi. Andmemudeli loomisel on taotletud võimalikult lihtsat ja ratsionaalset lahendust. Esialgselt loodud mudeli ligi 20 tabelit on koondunud lõpuks 15 tabeli peale. Samuti on mudeli loomisel arvestatud erinevate bioloogiliste aspektidega (nt. alternatiivne splaissing). Andmebaasis on hetkel 131 inimese metaboolset, signaali ülekande, rakuliste protsesside ja haiguslikku rada. Lisaks hulgaliselt keemilisi reaktsioone, ensüüme, keemilisi ühendeid ning geene. Põhirõhk on suunatud inimesega seotud bioloogilistele protsessidele. Suurimad probleemid, mis töö käigus tekkisid, olid erinevate bioloogiliste nüansside kaasamine mudelisse, adekvaatsete andmete töötlemine ja sisestamine andmebaasi ning kasutajasõbraliku veebiliidese loomine.

SUMMARY

A central problem in modern biological sciences is to understand the biochemical network of a cell. Scientists try to predict protein interaction networks that are responsible for certain processes in a cell. A grand challenge is a complete computer representation of the cell and the organism, which will enable computational prediction of higher-level complexity of cellular processes and organism behavior from genomic information.

Theoretical part of this work describes different metabolic pathway and protein-protein interaction databases. These databases are essential in the post-genome era for the analysis of genome. They describe genes and genomes, the predictive pathways, reactions, enzymes and metabolites of the organism. Visualization and analysis software afford to understand physiology of organism better. The key of a success in this field is a good knowledge of computer sciences, genomics and measurement technologies. A great challenge is understanding gene-regulatory logics and identifying biochemical networks. In the additional parts of this work, there are conclusive tables of wide-spread metabolic pathway databases including authors, main aims, volumes, models, advantages and tools.

The purpose of this work is to create a database for metabolic and signal transduction pathways including also data of chemical reactions, enzymes, genes and chemical compounds. When creating data model, I strove as rational and simple solution as possible. The final model consists of 15 tables with different biological aspects. The data model is adapted to add alternative splicing data. At the moment we have in the database information about 23683 genes, 4361 enzymes, 12642 chemical compounds and 131 human biological pathways. The most complicated problems were understanding and adding different biological aspects to the model, also working with data and creating web interface for the database.

VIITED

Aho AV and Ullmann JD (2000) Foundations of Computer Science (C edition).

Bader GD, Donaldson I, Wolting C, Ouellette BFF, Pawson T, Hogue CWV (2001) BIND – The Biomolecular Interaction Network Database. Nucleic Acids Research, Vol. 29, No. 1: 242-245

Bader GD and Hogue CWV (2000) BIND – a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics, Vol. 16, No. 5: 465-477

Becker MY and Rojas I (2001) A graph layout algorithm for drawing metabolic pathways. Bioinformatics, Vol. 17, No. 5: 461-467

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2004) GenBank: update. Nucleic Acids Research, Vol. 32: D23-D26

Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J *et al.* (2004) Ensembl. Nucleic Acids Research, Vol. 32:D468-D470

Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT (2003) MGD: The Mouse Genome Database. Nucleic Acids Research, Vol. 31:193-195

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research, Vol. 31:365-370

Costanzo MC, Crawford ME, Hirschman JE, Kranz JE, Olsen P, Robertson LS, Skrzypek MS, Braun BR, Hopkins KL, Kondu P, Lengieza C, Lew-Smith JE, Tillberg M and Garrels JI (2001) YPD™, PombePD™ and WormPD™:model organism volumes of the BioKnowledge™ library, an integrated resource for protein information. Nucleic Acids Research, Vol. 29:75-79

Dwight SS, Balakrishnan R, Christie KR, Costanzo MC, Dolinski K, Engel SR, **Feierbach B, Fisk DG, Hirschman J, Hong EL, Issel-Tarver L, Nash RS, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Weng S, Botstein D, Cherry JM** (2004) Saccharomyces genome database: underlying principles and organisation. Brief Bioinform. Mar,5(1):9-22

Goto S, Nishioka T and Kanehisa M (1998) LIGAND: chemical database for enzyme reactions. Bioinformatics, Vol. 14, No. 7: 591-599

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstroff P, Valencia A, Margalit H, Armstrong J, Bairoch A,

Cesareni G, Sherman D and Apweiler R (2004) IntAct: an open source molecular interaction database. Nucleic Acids Research, Vol. 32:D452-D455

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature, 415:180-183

Hodges PE, McKee AH, Davis BP, Payne WE, Garrels JI (1999) The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. Nucleic Acids Research, Vol. 27:69-73

Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y (2000) Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci U S A. 2000 Feb 1;97(3):1143-7

Joshi-Tope G, Gillespie M, Västrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu G, Matthews L, Lewis S, Birney E and Stein L (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Research, Vol. 33:D428-D432

Joshi-Tope G, Västrik I, Gopinathrao G, Matthews L, Schmidt E, Gillespie M, D'Eustachio P, Jassal B, Lewis S, Wu G, Birney E, Stein L (2003) The Genome Knowledgebase: A Resource for Biologists and Bioinformaticists. Cold Spring Harbor Symposia on Quantitative Biology. Volume LXVIII

Kanehisa M (1996) Toward pathway engineering: a new database of genetic and molecular pathways. Science & Technology Japan, No. 59: 34-38

Kanehisa M, Goto S, Kawashima S and Nakaya (2002) The KEGG databases at GenomeNet. Nucleic Acids Research, Vol. 30, No.1: 42-46

Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Research, Vol. 32, Database issue: D277-D280

Karp PD (2001) Pathway Databases: A Case Study in Computational Symbolic Theories. Science, Vol. 293: 2040-2044

Karp PD, Krummenacker M, Paley SM and Wagg J (1999) Integrated pathway/genome databases and their role in drug discovery. Trends in Biotechnology, Vol. 17(7): 275-281

Karp PD & Paley S (1994) Automated drawing of metabolic pathways. *Third International Conference on Bioinformatics and Genome Research*

Karp PD, Paley SM and Romero P (2002) The pathway tools software. Bioinformatics, Vol. 18, Suppl. 1:S1-S8

Karp PD, Riley M, Paley SM and Pellegrini-Toole A (2002) The MetaCyc Database. Nucleic Acids Research, Vol. 30, No. 1: 59-61

Karp PD, Riley M, Saier M, Paulsen IT, Paley SM and Pellegrini-Toole A (2000) The EcoCyc and MetaCyc databases. Nucleic Acids Research, Vol. 28, No. 1: 56-59

Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C and Gama-Castro S (2002) The EcoCyc Database. Nucleic Acids Research, Vol. 30, No. 1: 56-58

Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY and Karp PD (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Research, Vol. 32, Database issue: D438-D442

Krull M, Voss N, Choi C, Pistor S, Potapov A and Wingender E (2003) TRANSPATH: an integrated database on signal transduction and a tool for array analysis. Nucleic Acids Research, Vol. 31, No. 1. 97-100

Lemer C, Antezana E, Couche F, Fays F, Santolaria X, Janky R, Deville Y, Richelle J and Wodak SJ (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes. Nucleic Acids Research, Vol. 32, Database issue: D443-D448

Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schuller C, Stocker S, Weil B (2000) MIPS: a database for genomes and protein sequences. Nucleic Acids Research, Vol. 28: 37-40

Ogata H, Bono H, Fujibuchi W, Goto S, Kanehisa M (1996) Analysis of binary relations and hierarchies of enzymes in the metabolic pathways. Genome Informatics, Vol. 7: 128-136

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A *et al.* (2003) The Arabidopsis Information Resorce (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Research, Vol. 31(1):224

Rojdestvenski I (2003) Metabolic pathways in three dimensions. Bioinformatics, Vol. 19, no.18: 2436-2441

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Research, Vol. 32:D449-D451

Takahashi N and Smithies O (2004) Human genetics, animal models and computer simulations for studying hypertension. Trends in Genetics, Vol. 20, No. 3: 136-145

The FlyBase Consortium (2003) The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Research, Vol. 31:172-175

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nature Genetics, Vol. 25: 25-29

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. Genome Research, Vol. 11: 1425-1433

Westbrook J, Feng Z, Jain S, Bhat TN, Thanki N, Ravichandran V, Gilliland GL, Bluhm W, Weissig H, Greer DS *et al.* (2002) The Protein Data Bank: unifying the archive. Nucleic Acids Research 30:245-248

Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, Vinayaka CR, Zhang J, Barker WC (2003) The Protein Information Resource. Nucleic Acids Research 31(1):345-7

Xenarios I and Eisenberg D (2001) Protein interaction databases. Current Opinion in Biotechnology, Vol. 12: 334-339

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM and Eisenberg D (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Research, Vol. 30, No. 1: 303-305

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, Mitchell T, Miller P, Dean RA, Gerstein M, Snyder M (2001) Global analysis of protein activities using proteome chips. Science, 293:2101-2105

Lisa 1. Levinumate radade andmebaaside autorid, eesmärgid ja mahud.

Andmebaas	Autorid	Eesmärk	Mahud
MetaCyc Versioon 9.0	P.D. Karp S. Paley R. Caspi C. Fulcher B. Hopkinson P. Kaipa J. Pick	Metabolismi universumi kaardistamine	Rajad: 547 Rajad kommentaaridega: 300 Ensümaatilised reaktsioonid: 5046 Ensüümid: 2062 Ensüümid kommentaaridega: 1797 Geenid: 2133 Keemilised ühendid: 3945 Organismid: 341 Viited: 5468
EcoCyc Versioon 9.0	P.D. Karp I. Keseler A. Shearer S. Paley J. Pick	*E.coli* molekulaarse kataloogi kirjeldamine, sh. iga molekulaarse osa funktsioonid	Rajad: 183 Reaktsioonid: 3634 Ensüümid: 1148 Transporterid: 196 Valkude kommentaarid: 3609 Geenid: 4476 Transkr.ühikud: 983 Viited: 9873
aMAZE 2004.a.	S.J.Wodak J. Richelle C. Lemer J. van Helden	Infoallikas interaktsioonide ja protsesside kohta, võimaldades samal ajal kompleksanalüüse	EC number: 4219 Ekspressioon: 12830 Geen: 12689 Organism: 3 Polüpeptiid: 61249 Protsess: 106 Reaktsioon: 5281 Spets. ühendid: 10464 Transkr.regulatsioon: 583 Metaboolsed rajad: 106 Sign. transd.-i transformatsioonid: 338 Sign. transd.-i kontroll: 298 Sign. transd.-i rajad (*S.cerevisiae*): 18

KEGG: PATHWAY GENES SSDB KO LIGAND = COMPOUND + GLYCAN + REACTION + ENZYME Versioon 34.0	M. Kanehisa S. Goto	Esitada arvutis kogu informatsioon biomolekulide radadest	GENES - valk: 891942 GENES - ensüüm: 166318 GENES - rada: 156598 COMPOUND sisendeid: 12817, neist keem. valemiga:12118 mol. struktuuriga: 12111 GLYCAN sisendeid: 11020, neist klassiga: 7095 REACTION sisendeid: 6406, neist nimega: 4652 ENZYME sisendeid: 4361, neist süstemaatilise nimega: 3475
Reactome Versioon 13	S. Lewis E. Birney I. Västrik L. Stein G. Joshi-Tope	Inimese radade ja reaktsioonide allikas	Mitootiline rakutsükkel: 157 molekuli Rakutsükli kontrollpunktid: 73 molekuli DNA reparatsioon: 164 molekuli DNA replikatsioon: 106 molekuli Insuliini retseptori aktivatsioon: 24 molekuli Lipiidide metabolism: 119 molekuli Aminohapete metabolism: 199 molekuli Glükoosi jt. suhkrute ning etanooli metabolism: 147 molekuli mRNA protsessimine: 153 molekuli Nukleotiidide metabolism: 318 molekuli Püruvaadi oksüdatiivne dekarboksüleerimine: 49 molekuli Transkriptsioon: 109 molekuli Translatsioon: 124 molekuli

Lisa 2. Levinumate radade andmebaaside mudelid, eelised ja tööriistad.

Andmebaas	Mudel	Eelised	Tööriistad
MetaCyc	Objekt-orienteeritud	Ulatuslikud kommentaarid radade ja ensüümide kohta Viited kirjanduslikele allikatele Iga liigi jaoks oma rada Rajad märgistatud infoga, millisel liigil antud rada eksperimentaalselt kindlaks tehtud Andmed spets. ensüümide omaduste kohta eri liikidest	PathwayTools - päringud, visualiseerimine: 1) PathoLogic 2) Pathway/Genome Navigator 3) Pathway/Genome Editors
aMAZE	Objekt-orienteeritud	Selge erinevus mol. objektide ja sündmuste vahel Füüsikalised ja funktsionaalsed interaktsioonid geenide ja geeniproduktide vahel Komplekssete funktsioonide ja ruumiliste asukohtade käsitlemine Suure hulga protsesside kirjeldamine Erinevad klassifikatsiooniskeemid	aMAZE LightBench veebiliides aMAZE WorkBench: andmete laadimine ja pakkimine, modifikatsioonid-annotatsioonid, visualiseerimine ja analüüs
KEGG: PATHWAY GENES SSDB KO LIGAND = COMPOUND + GLYCAN + REACTION + ENZYME	Relatsiooniline	Ensüümi funktsionaalsete ülesannete uurimine Geeniproduktide funktsiooni ennustamine Met.radade visualiseerimine liikidevaheliseks võrdluseks Radade analüüs ja neis olulist rolli mängivad ensüümid Millised ensümaatilised sammud ennustatavalt toimuvad antud rajal mitmetes sekveneeritud genoomides Tundmata geeni produkti biol.funkts.-de identifitseerimine Radade konstrueerimine Erinevate liikide geenide ja genoomide võrdlev analüüs Erinevates rakkudes geenikspressiooni simuleerimine, analüüs	Objektide värvimine radade kaardil Mitmed analüüsitööriistad Lühima tee arvutamine
Reactome	Relatsiooniline	Inimese bioloogilistele radadele orienteeritud andmebaas	Pathfinder – lühim tee kahe molekuli vahel Skypainter – visualiseerimine