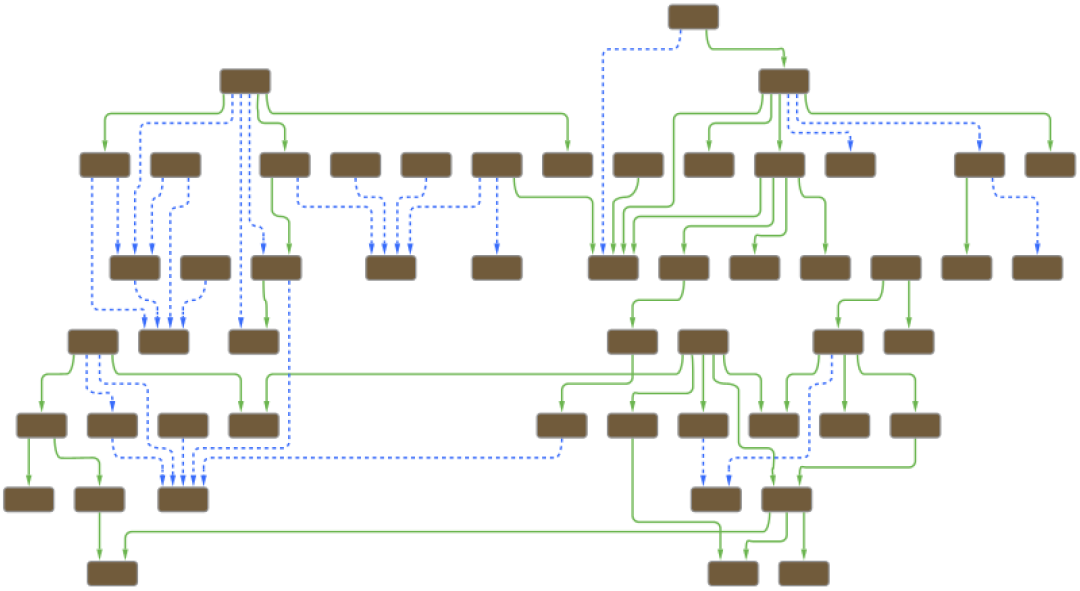
CHALMERS



UTFORMANDE OCH ANALYS AV SLÄKTTRÄDET ÖVER GEMs

Lokalisering, insamling, analys och kategorisering av kopplingar mellan GEMs utifrån modellåteranvändning, i samband med en kvalitativ studie

Kandidatarbete inom Institution för Biologi och Bioteknik
BBTX01-21-04

Elin H. Berntsson, Kennet Lindquist, Yi Le
Lisa-Marie Witte, Gaston Sandstig, Jewahri I. Yousuf

Avdelningen för Infrastrukturer inom Institutionen för Biologi och Bioteknik

CHALMERS TEKNISKA HÖGSKOLA
Göteborg, Sverige 2021
www.chalmers.se

KANDIDATARBETE 2021

UTFORMANDE OCH ANALYS AV SLÄKTTRÄDET ÖVER GEMs

Lokalisering, insamling, analys och kategorisering av kopplingar mellan GEMs utifrån modellåteranvändning, i samband med en kvalitativ studie

Elin H. Berntsson, Kennet Lindquist, Yi Le
Lisa-Marie Witte, Gaston Sandstig, Jewahri I. Yousuf

CHALMERS

Institutionen för Biologi och Bioteknik
Avdelningen för Infrastrukturer
Chalmers tekniska högskola
Göteborg, Sverige 2021

UTFORMANDE OCH ANALYS AV SLÄKTTRÄDET ÖVER GEMs
Lokalisering, insamling, analys och kategorisering av kopplingar mellan GEMs utifrån modellåteranvändning, i samband med en kvalitativ studie
Elin H. Berntsson, Kennet Lindquist, Yi Le
Lisa-Marie Witte, Gaston Sandstig, Jewahri I. Yousuf

Huvudhandledare: Jonathan Robinson, Avdelningen för Infrastrukturer, Institutionen för Biologi och Bioteknik
Förslagsställare/kontaktperson: Mihail Anton, Avdelningen för Infrastrukturer, Institutionen för Biologi och Bioteknik
Övriga handledare: Hao Wang, Avdelningen för Infrastrukturer, Institutionen för Biologi och Bioteknik
Examinator: Claes Niklasson, Institutionen för Kemi och Kemiteknik

Kandidatarbete 2021
Institutionen för Biologi och Bioteknik
Avdelningen för Infrastrukturer
Chalmers Tekniska Högskola
SE-412 96 Göteborg
Telefon +46 31 772 1000

Länk till projektets GitHub-sida: https://github.com/MetabolicAtlas/tree-of-GEMs

Cover: Visualisering av återanvändning mellan metaboliska modeller, klassificerade som helt eller delvis direkta med grön respektive blå färg.

Typeset in LaTeX

Acknowledgement

We express deep and sincere gratitude to our supervisors Mihail Anton, Hao Wang and Jonathan Robinson, whose guidance, suggestions and detailed constructive criticism have contributed to the development of the thesis. We would also like to extend our gratitude for providing all required resources to increase our knowledge to achieve the best results.

We also want to thank all five researchers that we were able to interview during the course of this project. Their knowledge provided a great insight into the field of systems biology and specifically GEMs, and we are very grateful for their time and effort in being a part of this project.

Elin H. Berntsson, Kennet Lindquist, Yi Le,
Lisa-Marie Witte, Gaston Sandstig, Jewahri I. Yousuf

Gothenburg, May 2021

BUILDING AND ANALYSING THE TREE OF GEMS
Locating, collecting and organizing the connections between GEMs, based on model reuse, in conjunction with a qualitative study
Elin H. Berntsson, Kennet Lindquist, Yi Le
Lisa-Marie Witte, Gaston Sandstig, Jewahri I. Yousuf

Division of Infrastructure
Department of Biology and Biotechnology
Chalmers University of Technology

Abstract

Genome-Scale Metabolic Models (GEMs) describe reactions in the metabolic system of an organism based on its genes. These models are simulated for a variety of different applications, such as research on diseases or evolutionary processes. Since 1999, when the first GEM was published, the field has grown and several models have been constructed for a range of different organisms. The purpose of this work is to make the reuse and development of GEMs more accessible, transparent, and functional to the scientists and other parties within the field, also in accordance to the FAIR-principles. The aim of this project is to gather and analyse data of 182 GEMs, where the annotation ”model reuse” is enriched with a newly defined set of annotation parameters. The final result provides a compact and concise visualization of the information. Annotation parameters such as model name, publication year, and model inheritance were added to a dataset consisting of 107 articles with 182 models. The process of annotating the dataset was partially automated with the use of MATLAB. The result of the annotation is a dataset in the form of a JSON-file, which is used to visualize a family tree of GEMs, where the data is sorted by publication year and a classification of organisms. A qualitative study was carried out as a complimentary method, by interviewing five researchers within the field. The thoughts and opinions that were collected throughout the interviews inspired both the current questions that are answered in the project as well as ideas for future projects. A number of ideas, such as integration of version control and ID of model components could contribute to an increased in ease of accessibility and reusability of GEMs. The dataset and its visualization opens up the possibility for increased transparency in the field of Genome-Scale Metabolic Models, which in turn benefits and promotes a continued development and standardization of GEMs.

Keywords: Genome-Scale Metabolic Model, GEM, model reuse, model inheritance, constraint-based model

Sammanfattning

Genome-Scale Metabolic Models (GEMs) beskriver reaktionerna i en organisms metaboliska system utifrån dess gener. Modellerna simuleras för många olika applikationer, såsom forskning på sjukdomar eller evolutionära processer. Sedan 1999, då den första GEM publicerades, har området växt och flera modeller har konstruerats för en rad olika organismer. Syftet med projektet är att göra återanvändningen och utvecklingen av GEMs mer tillgänglig, transparent och funktionell för forskare och andra parter inom området. Projektets utformning skedde enligt FAIR-principerna, som är en viktig aspekt inom systembiologi-branschen. Målet med detta projekt är att samla och analysera data av 182 GEMs, där annoteringen ”modellåteranvändning” berikas med ett nydefinierat set av annoteringsparametrar. Den slutgiltiga produkten ger en kompakt och koncis visualisering av denna information. Annoteringsparametrar såsom modellnamn, publiceringsår och modellarv adderades i ett dataset som utgörs av 107 artiklar med tillhörande 182 modeller. Med hjälp av MATLAB kunde processen för annotering av datasetet delvis automatiseras. Resultatet av annoteringen är ett dataset i JSON-format, som användes för att visualisera ett släktträd över GEMs, där datan är sorterad efter både publiceringsår och en klassificering av organismer. En kvalitativ studie utfördes som en komplementär metod, genom att intervjua fem forskare som har GEMs som huvudforskningsområde. Åsikter och tankar samlades under intervjuerna och ansågs vara inspirerande för både de aktuella frågorna som besvaras i projektet och framtida projekt. En del idéer, såsom integrering av versionshantering och ID av modellkomponenter skulle kunna bidra till att öka lättillgängligheten och återanvändbarheten av GEMs. Datasetet och dess visualisering öppnar möjligheter för ett mer transparent GEM-område som i sin tur gynnar och främjar en fortsatt utveckling och standardisering av GEMs.

Innehåll

Figurer

Tabeller

1
Inledning

I det här arbetet står Genome-Scale Metabolic Models (GEMs) och återanvändningen av dem i fokus. När branschen inom GEMs expanderar blir det allt mer nödvändigt att utveckla tidigare kunskaper, och iterera på eller modifiera tidigare modeller. Således blir modellers återanvändning central för området. I detta arbete gestaltas återanvändning som arv mellan modeller, eftersom hela eller delar av tidigare skapade GEMs ”ärvs” av efterföljande rekonstruerade sådana. Information om återanvändning finns disponibel i de artiklar som publiceras samtidigt som modeller tillgängliggörs, men återges ej på ett standardiserat vis och finns inte samlad utanför artiklarna. Det försvårar arbetet för de forskare och andra parter som arbetar med GEMs i någon utsträckning, av detta skäl finns det ett värde i att göra denna information åtkomlig och sökbar.

Projektet har som syfte att via konstruktionen av ett dataset gällande de kopplingar som modellåteranvändningen utgör, kunna gynna forskare i den fortsatta utvecklingen av GEMs. Projektet har även som målsättning att samla, analysera och annotera data i enlighet med FAIR-principerna för att uppnå en mer öppen och transparent forskning inom området. För att uppnå både mål och principerna implementeras metodik som till viss del är manuell, men också automatiserad med hjälp av datorprogrammet MATLAB. För att optimera projektet och dess resultat utförs även en kvalitativ studie, där ett flertal forskare inom GEM-området intervjuas för att kunna identifiera de viktigaste problemen. Målet är även att i slutskedet skapa en visualisering över modellarv som förhoppningsvis blir grunden till ett framtida större släktträd över GEMs.

1.1 Bakgrund

Metabolism, på en stor och liten skala, spelar en vital roll för alla celler och organismer. Hur det metaboliska nätverket fungerar i samspel med resten av kroppen är därför en av nycklarna i förståelsen av oss själva och andra organismer. Men nätverket är inte fritt tillgängligt i en levande organism, då det både är internt och föränderligt. Olika delar av nätverket står också i kontakt och interagerar med varandra genom till exempel feedback-mekanismer. Det blir därför komplicerat att säga vilka effekter en viss sjukdom har på systemet och vice versa vilka systemtillstånd som kan kopplas till bland annat stress eller välmående.

Ett sätt att studera nätverkets rådande situation är genom dess produkter, det vill säga metaboliter. Ett exempel är NMR-fingeravtrycket från ett urinprov för

att diagnostisera cancer, MSUD (Maple syrup urine disease) eller diabetes [1]. Men metaboliter representerar bara resultatet av systemet, och inte de högst variabla vägarna en kol-, kväve- eller syreatom tagit igenom det. Varje fingeravtryck kan alltså potentiellt representera en stor mängd olika systemtillstånd [2]. En mer nyanserad bild ges med hjälp av ”stable isotope probing” (SIP) där färdtiden för en isotop genom systemet analyseras. Kombinerat med en modell av det metaboliska nätverket kan man då kvantifiera aktiviteten av olika reaktioner och få stor inblick i systemets tillstånd [3]. I det här projektet är det denna metaboliska modell, även kallat GEM, som står i centrum.

1.1.1 Genome-Scale Metabolic Model

Den första Genome-Scale Metabolic Model (metabolisk modell i genomskala, eller GEM) konstruerades för bakterien *Haemophilus influenzae* 1999 och sedan dess har många fler modeller utformats. GEMs har konstruerats inte bara för bakterier, utan också för arkéer och eukaryoter [4]. Dessa används idag inom många olika studier med diverse fokus, till exempel de metaboliska sjukdomarna dysbios, diabetes mellitus och fettlever (leversteatos). Men GEMs användningsområde sträcker sig även utanför studier på sjukdomar av metabolisk karaktär, bland annat cancerstudier [5]. Modeller över metaboliska nätverk nyttjas även i utvecklandet av läkemedelsmål rörande viktiga patogener; studerandet av evolutionära processer eller för att tolka vissa experimentella resultat [6].

Metaboliska modeller av den här sorten är heller inte begränsade till sjukdomsstudier och forskning om exempelvis människans metabolism, utan appliceras även inom industrin. Modeller av *Saccharomyces cerevisiae* eller *Escherichia coli*, och den kunskap som erhålls av dessa, är exempelvis mycket tillämpbara inom produktion av biobränslen och relaterade ämnen såsom ett ökat produktutbyte [5]. GEMs applikationsområde är alltså brett och deras vidareutveckling kan därför vara till grund för ökad förståelse inom biovetenskap och ny bioteknologi [4].

1.1.1.1 Utvecklingen av databaser för GEMs

När de första GEMs konstruerades tillgängliggjordes de genom redan etablerade modelleringsdatabaser. Allt eftersom intresset att konstruera GEMs har växt har också nya databaser skapats för att fokusera exklusivt på metaboliska modeller. I enlighet med detta är det av betydelse att förstå hur pass tilltagande utvecklingen av nya databaser är och när detta ökade intresse tog fart.

Abraham m.fl. [7] identifierar i sin artikel från 2017 de 42 mest använda och populära databaser som inkluderar metaboliska modeller. Ett flertal av databaserna är nu tillsynes nedlagda. Av resterande 27 databaser som fortfarande används idag grundades den första 1987 och ytterligare 4 resurser etablerades innan år 2000 [8]–[12]. Nästan 50% av databaserna som fortfarande finns åtkomliga idag och där information forfarande finns till förfogande, grundades 2010 eller senare [13]–[34].

I samband med att meta-databasen ”Pathguide: The Pathway Resource List” etablerades 2006 publicerades även en tillhörande artikel [35]. Här identifieras de dåvarande

tillgängliga databaserna för biologiska nätverk, vilka också var samlade i listan som Pathguide omfattade. Vid start innefattade den här listan över 190 databaser, av vilka 44 stycken innefattade metaboliska modeller. Denna guide över databaser används fortfarande idag och de 44 första resurserna har nu ökat till 166 stycken. Det är en ökning på nästan 280% sedan 2006.

Med ett ökande antal databaser kommer även ett ökat antal sätt att konstruera och utforma modeller, vilket har att göra med de olika delarna en modell innefattar. Följande avsnitt klargör detta mer utförligt.

1.1.1.2 Databaser och modelleringsverktyg för GEMs

Processen att skapa modeller över metaboliska nätverk är både tidskrävande och mycket komplex, vilket har drivit på utvecklingen av hjälpmedel inom området. I nuläget finns en stor variation av metoder och modelleringsverktyg tillgängliga, såsom CarveMe [36], Path2Models [37] och AGORA [38]. Således kan modeller nu utformas manuellt eller på ett mer automatiserat sätt. Samtliga verktyg har både för- och nackdelar, exempelvis hur pass automatiserad processen är, om användaren tillåts göra förbättringar manuellt eller om verktyget är väl standardiserat [6].

Vidare finns det, som redan nämnts, ett flertal databaser som omfattar alla modeller som hittills har publicerats. Exempel på databaser är BioModels [39], Biochemical-Genomic and Genetic knowledgebase (BiGG) Models [40] och Kyoto Encyclo-pedia of Genes and Genomes (KEGG) [41]. Dessa inbegriper antingen en mängd olika modeller, GEMs bland annat, eller så står endast metaboliska modeller i fokus.

BioModels dataset av modeller uppgår exempelvis för tillfället till 2572 modeller, varav 162 av dessa är begränsnings-baserade (constraint-based) där GEMs oftast ingår. Dessa 162 modeller med 87 tillhörande artiklar har olika strukturer, vilket syftar på vilket sorts verktyg som använts eller om konstruktionen av modellen har utförts manuellt. Emellertid erbjuder också databasen vissa specifikt tillägnade söksidor där fler modeller av olika typ finns tillgängliga [42]. Detta är ett resultat av olika projekt, bland annat ”Path2Models project” där cirka 140 000 modeller, varav cirka 10 000 GEMs, för olika organismer automatiskt genererades utifrån ett flertal databasers resurser [37]. KEGG, Mitelman Database och MetaCyc är exempel på databaser som medverkade [43].

1.1.1.3 Konstruktion och struktur av GEMs

Vid konstruktionen av en GEM associeras ett ID (identifierare) med de olika delarna av modellen, såsom metaboliter, reaktioner, cellkompartement eller gener [44]. Exempel på ID som är centrala för det här arbetet är metabolit- och reaktions-ID. De kan vara unika för modellen, men i en majoritet av fallen är de direkt kopplade till en specifik databas [45]. Vilken databas som används bestäms av modelleringsverktyget, till exempel får modeller skapade med programmet RAVEN sina identifierare från KEGG [46], [47]. I nuläget finns emellertid ingen standard för identifiering, så även grundläggande modelldelar såsom ATP eller DNA har ofta olika ID i de olika databaserna. Nästan 60% av GEMs saknar standardiserade metabolit-ID redogjorde

Ravikrishnan och Raman för i en artikel publicerad 2015 [48]. Saknaden av konsensus och en fortsatt utveckling av nya verktyg och databaser skapar en stor heterogenitet inom identifiering [45], vilket komplicerar både jämförelsen av modeller och försök att kombinera dem [44].

1.1.1.4 Problematik och förslag på lösningar gällande standardisering av GEMs

Just standardisering av GEMs och dess komponenter är ett centralt problem inom systembiologi-branschen, vilket är en effekt av olika orsaker som påverkar konstruktionen av modeller och även simuleringen av dessa [49], det vill säga hur väl modellen speglar det verkliga metaboliska systemet. Exempel på faktorer som berör standardisering är i vilket filformat modellen lagras och delas genom, namngivning av ID och hur modellen kvalitetstestas [49].

Filformatet som anses vara standard för modeller inom området är SBML (Systems Biology Markup Language) [50] och innehåller kod för alla komponenter en modell består av, såsom metaboliter och gener. Koden innehåller också tillhörande begränsningar som är kopplade till komponenterna, exempelvis specifika cellkompartement eller reaktioner. När en GEM skapas och sedan sparas i SBML-format sker en automatisk standardisering [49], men i 36% av fallen används andra format [48]. Således blir detta en påverkande faktor, vilket tillsammans med icke-standardiserade ID förstärker problematiken och hör även ihop med kvalitetstester av GEMs. Modeller bör analyseras så att namngivning av ID har skett korrekt och hur väl modellen stämmer överens med biologiska system [51].

Utifrån de faktorer som kort förklarats ovan och som mer djupgående diskuteras av bland annat Carey m.fl. [49] finns vissa förslag på lösningar. Incitament kan skapas genom att finansiera forskningsprojekt riktade åt standardiserade metoder och verktyg. Dock har denna lösning vissa generella problem då finansiering av forskning utvärderas på ett kortsiktigt plan, men resultaten av projekten sker långsiktigt. Ytterligare en lösning som presenteras är att modellåteranvändning bör främjas och uppmuntras, eftersom det lägger grund för en öppen kommunikation mellan alla parter som arbetar med GEMs [49]. Rekonstruktionsjamborees, där forskare träffas för att kombinera sina kunskaper och modellprototyper till en sammanställd modell, kan ses som ett exempel av en sådan ökad kommunikation [52].

1.1.1.5 Återanvändning av GEMs

En viktig faktor av rekonstruktionen av GEMs är återanvändningen av äldre modeller. Modellåteranvändning kan ske på många sätt, såsom att den nya modellen baseras på den äldre men med nya inslag och modelldelar. Eller när den rekonstruerade modellen integrerar delar av en eller flera tidigare skapade modeller för att utforma en ny, mer enhetlig sådan. Dessutom finns möjligheten att kombinera äldre modeller i syfte att rekonstruera en GEM som är mer omfattande [44]. Därutöver finns ännu fler alternativ för hur en modell kan konstrueras och modelleringsverktygen är många.

Begreppet ”annotering” används i detta projekt och kan förklaras som specifik information kopplat till modellerna. När denna information avser ”modellåteranvändning” som annotering finns även flera annoteringsparametrar, såsom publiceringsår och art. Andra viktiga parametrar som kan inkorporeras är modellarv (model inheritance), modellversion, ID och filformat.

Just modellarv och versioner av modeller är viktiga parametrar när modellåteranvändning står i fokus. Modellarv hör främst till när äldre publicerade modeller används som grund eller inspiration för nya rekonstruerade GEMs, som beskrivet ovan. Här publiceras en för tillfället färdigställd GEM och ofta tillsammans med en vetenskaplig artikel. Versionshantering av modeller skiljer sig i det avseendet att konstruktionen av modellen är mer testdriven [51].

För varje version kan förbättringar göras och essentiella modelldelar integreras av en eller flera medverkande. Mellan varje version sparas förändringarna som historik, vilket gör det möjligt att utgå från olika versioner när en ny modell skapas. Om den nyaskapade modellen ej visar sig vara tillämplig, finns valet att återgå till en av de tidigare versionerna. Förbättringar mellan olika förgreningar av modellen kan även kombineras till samma modell förutsatt att de inte strider mot varandra. För att på ett lättillgängligt och transparent sätt hantera versionshantering kan den ske på offentliga repositorium (datalager) såsom GitHub och SourceForge.

Det är många faktorer som spelar in hur en GEM exempelvis ska konstrueras, ha för struktur eller vart den ska publiceras. Dessutom ökar antalet GEMs konstant och det blir svårare att välja en bra utgångspunkt för studier och eventuella modeller. Alla dessa påverkande faktorer leder vidare till frågan om processen att fortsätta utveckla modellåteranvändning kan förenklas och göras mer gynnsam för forskare som arbetar inom området idag.

1.2 Problemformulering

Eftersom värdet och intresset av att skapa GEMs har växt i stadig takt har också antalet modelleringsverktyg och databaser som inbegriper dem ökat. Modeller skapas på nytt eller rekonstrueras för många olika organismer genom diverse metoder, och modellåteranvändning spelar en stor roll. Men heterogeniteten inom ID av modelldelar och saknaden av fullständigt utvecklade annoteringsparametrar har resulterat i ett forskningsområde med avsaknad av FAIR-principerna. ”FAIR” står för Findable, Accessible, Interoperable, Reusable (sökbar, tillgänglig, interoperabel, återanvändbar).

Som återanvändningen och utvecklingen av GEMs ser ut idag finns ingen konsensus i hur information kring dem ska tillgängliggöras. För tillfället hittas denna fakta, exempelvis kopplingar mellan olika GEMs, i de artiklar som publiceras i samband med en modell. Som forskare eller andra aktiva inom området krävs en egen, ofta tidskrävande och arbetsam, insats om man vill hitta denna fakta för egen användning.

Med andra ord är återanvändning av GEMs i större skala arbetskrävande, men

viktig av den anledning att tidigare kunskap kan fortsätta byggas på. Transparent återanvändning av GEMs kan underlätta för fortsatt utveckling av området och gynna ett bättre samarbete inom forskningen. Ytterligare anledning till att återanvända modeller är för att, som beskrivs av Carey m.fl., uppmuntra standardisering av modellkonstruktionen som den ser ut idag. Frågeställningen blir följaktligen om återanvändningen av GEMs på något sätt kan göras mer transparent och lättillgänglig och vilka möjliga frågor detta skulle kunna besvara.

1.3 Syfte och mål

Projektet syftar till att underlätta och göra utvecklingen av GEMs mer lättillgänglig för forskare och andra parter inom systembiologi-branschen i enlighet med FAIR-principerna, vilket även skulle gynna standardiseringen av modeller. För att uppnå en mer lätthanterlig och transparent återanvändning av GEMs har det här projektet som mål att samla och analysera data om modellåteranvändning. Samtidigt inkluderas också åsikter och förslag på fokusområden enligt olika forskare för att identifiera de mest angelägna problemen inom branschen. Detta bidrar även till att kunna anpassa befintliga modeller efter forskares behov och på detta sätt gynna vidareutvecklingen av GEMs i framtiden.

Målsättningen är att berika de 162 befintliga GEMs (exkluderat modeller från Path2Models projektet) som finns på databasen BioModels med nya annoteringsparametrar som tydliggör modellåteranvändning. Målet är även att manuellt lägga till väl citerade modeller för att utöka datasetet, vilket därmed kommer ge en mer enhetlig grund för analys. Slutprodukten skall kunna ge en kompakt och lätthanterlig visualisering av information som beskriver återanvändning av GEMs, modellarv, samt eventuellt versionshantering, vilket skulle kunna gynna forskare i sin tur.

Projektet har även som mål att genom behandling av data och utvärdering av befintlig information kunna besvara följande frågor:

- Hur stor andel av modeller har någon koppling?
- Finns det modeller som återanvänds mer än andra, och indikerar det en högre kvalitet?
- Hur ofta uppdateras modeller?
- Hur ser trenden ut för antalet publicerade modeller per år?
- Har det skett en förändring i antalet modeller som återanvänds per rekonstruerad modell genom åren?

1.4 Etik

1.4.1 Annotering av modeller i samband med FAIR-principerna

Projektets ändamål kan jämföras med FAIR-principerna som är vetenskapliga riktlinjer för datahantering inom vetenskaplig forskning. Syftet med FAIR-principerna och huvudfrågan av projektet är att göra forskningsdata sökbara, tillgängliga, interoperabela och återanvändbara [53]. Genom att berika modellers metadata och

tydligare beskriva återanvändningen av dessa blir det enklare för en bredare målgrupp att förstå kopplingar mellan modeller och fortsätta utvecklingen av dem. Datasetvisualiseringen som skapas i det här projektet kan därmed vidareanvändas av forskare inom området och på så sätt underlätta deras arbete.

1.4.2 Beaktande av intervjudeltagarnas integritet och hantering av personlig information

Intervjuerna och undersökningen har som syfte att genomföras med alla deltagares godkännande, vilket förmedlas till dem i förväg. Om syfte eller informationshanteringen ändras under projektarbetets gång kontaktas de intervjuade forskarna för nytt godkännande av användningen av insamlad information. Autonomin och integriteten för varje respondent som bidragit till slutrapporten respekteras och åtkomst till personlig data begränsas. Ingen personlig information som identifierar individerna avslöjas i den slutliga rapporten om de inte själva begär det. Vid redogörandet av resultatet från intervjuerna benämns deltagarna med bokstäverna A, B, C, D, E, som tilldelats slumpvis, istället för deras namn.

1.4.3 Hantering av det innehåll som framkommer under intervjuerna

Respondenterna informeras om att intervjuerna har som avsikt att spelas in (endast ljudet) i ändamål att sammanfatta och granska svaren på ett korrekt och angeläget sätt. Forskarnas rekommendationer, med avseende på deras akademiska bakgrund, om tydligare annoteringsparametrar kommer beaktas och integreras i utformningen av projektet. Vid summering av intervjudeltagarnas åsikter och svar kommer direkta citat att antecknas och tolkas opartiskt efter bästa möjliga förmåga. Efter projektets gång och den slutgiltiga rapporten färdigställts kommer ljudinspelningar raderas tillsammans med respektive summering. Vissa citat eller tolkade svar anges i avsnitt 3.1.

1.5 Avgränsningar

1.5.1 Både automatiserad och manuell datainsamling implementerades

Datainsamling- och bearbetning är en avgörande del i detta projekt. För att underlätta och tidsmässigt effektivisera hela processen från datainsamling i början av arbetet till visualisering i slutet, valdes att konstruera en rad olika MATLAB-koder. Projekttiden kunde användas effektivt genom att automatisera exempelvis nedladdning av modeller från olika databaser eller uppladdning av klassificeringar till versionshanteringsverktyget GitHub.

Förbättringar i processen kunde implementeras med jämna mellanrum, utan att att det krävde mycket tid eller resurser. Dessa MATLAB-koder kompletterades med manuell datahantering, eftersom datan efter insamlingen fortfarande behövde tolkas.

På grund av projektets tidsbegränsning och projektdeltagarnas avsaknad av kunskap inom ämnet valdes det att inte implementera någon form av maskininlärning för att utföra dessa tolkningssteg. Fokus lades istället på att hitta en gemensam grund om vilka tolkningskriterier som kunde användas vid den manuella datahanteringen.

I början av projektet planerades att undersöka och annotera cirka 1000 modeller. Efter annotering av 179 modeller konstaterades det att större delen av det då aktuella datasetet ej innehöll modeller som kunde klassificeras som GEMs. I brist på kvarvarande tid av projektet och en kod anpassad till inhämtning av dataset från BioModels [39] reducerades antalet modeller till endast 162 stycken, men nu endast GEMs. Detta åstadkoms genom att begränsa sökningen till begränsningsbaserade modeller vid nedladdning av metadata, vilket är beskrivet i avsnitt 2.2.

I slutet då viss tid gick att tillgå kompletterades även det initiala datasetet med ytterligare modeller. Motivet till detta var att på så sätt säkerställa ett mer informativt släktträd av GEMs. För utförligare beskrivning av urvalet för dessa manuellt kompletterade modeller, se avsnitt 2.2. Här bestämdes emellertid att begränsa hur många modeller som kunde läggas till i efterhand, sådant att projektets tidsbegränsning kunde hållas i åtanke. Antalet GEMs som det fanns tid med att inkludera manuellt uppgick till 20 stycken.

1.5.2 Vissa annoteringsparametrar valdes bort

Det bestämdes att inte inkludera modellversion som aktiv annoteringsparameter i detta projekt. Huvudsakligen på grund av projektets tidsbegränsning, men även faktumet att modellversioner inte tillämpas av alla forskare som skapar GEMs förr och idag, vilket försvårar analysen. Däremot diskuteras användbarheten och tyngden av modellversioner i avsnitt 3.2.2.1.

Metabolit- och reaktions-ID är annoteringsparametrar som nämndes av nästan alla forskare under intervjuerna. Dessa parametrar diskuteras i avsnitt 3.2.2.2, men tillämpas inte i projektet, av samma anledning som för modellversion.

Det finns ytterligare funderingar och idéer på flera olika annoteringar och annoteringsparametrar. I avsnitt 3.2.3 ges därför även rekommendationer på sådana som skulle kunna användas i framtida forskningsprojekt.

1.5.3 Tidsbegränsningen påverkade intervjuhanteringen

Det intervjuades totalt fem personer i samband med detta projekt. Antalet påverkades av projektets utformning och den resulterande tidsbegränsningen. Fyra av fem respondenter är anställda på Chalmers, vilket är en följd av att forskare valdes till den kvalitativa studien utifrån handledarnas rekommendationer.

Intervjuerna spelades in, men det bestämdes att inte transkribera intervjuer på grund av projektets tidsramar. Däremot skrevs det sammanfattningar efter varje intervju där den mest väsentliga informationen klargjordes och delades sedan med gruppen.

2
Metod

2.1 Annoteringsparametrar för att beskriva modellåteranvändning

Projektets datasetannotering skedde i två vändor, paradigm 1 och 2. I båda fallen samverkade projektdeltagarna med kod för att analysera modellerna. Det första fallet omfattade Biomodels kurerade modeller medan det andra omfattade de begränsningsbaserade modellerna, runt 1000 respektive 100 till antal.

För att beskriva modellarv och återanvändning fastställdes fyra parametrar som varje modell annoterades med: kopplingar, modell-ID, taxon och publiceringsår. De första två ansågs essentiella för att beskriva annoteringen, medan resten ökar förståelsen och nyanserar informationen. Valet av parametrar grundades även i hur lättillgängliga de var. **Kopplingar** redogör för vilka modeller som återanvänds i den annoterade modellen, *i.e.* vilka parent-modeller som tillhör en child-modell. Parametern pekar alltså uppströms. Där återges om kopplingen är delvis eller helt direkt, ett utdrag ur parent-modellens artikel som motiverar kopplingen (med direktlänk till artikeln) och slutligen parent-modellens PubMed ID (PMID).

Om kopplingen är helt direkt har child-modellen byggd på sin parent, eller så är den en kombination av flera parent-modeller. Vid en delvis direkt koppling inkorpererar child-modellen endast delar av dess parent-modeller. I paradigm 1 fanns även en tredje sorts koppling, indirekt, som beskrev fenomen som verifiering baserad på andra modeller. Eftersom det inte beskrev arvet eller återanvändningen av modellernas innehåll förkastades klassificeringen i paradigm 2. **Modell-ID** inkluderar ett PMID, DOI samt Biomodels submission-ID. Modeller med samma tillhörande artikel är därför grupperade som en modell, alltså kan en modell ha flera submission-ID. Slutgiltligen annoterades **Taxon** genom ett trivialnamn, ett vetenskapligt namn samt med NCBIs taxonomi-ID [54][55].

I paradigm 1 var data spridd över ett Exceldokument och i MATLAB-format på projektets repositorium. I paradigm 2 skedde datahanteringen lokalt programatiskt i MATLAB där annoteringarna ligger i en hierarkisk struktur. Strukturen konverterades sedan till JSON-format innan den laddades upp till projektets repositorium. På sådant vis möjliggjordes versionshantering av datasetet och högre interoperabilitet.

2.2 Annotering av datasetet

I projektet användes Biomodels som utgångspunkt av datasetet. All tillhörande metadata hämtades ned genom MATLAB via API [56], [57]. Sedan extraherades modell-ID och publiceringsår från datan. Filtreringen av modellerna itererades för att exkludera icke-GEMs, och är anledningen till paradigmbytet.

I paradigm 1 kodades ett program för att ge potentiella kopplingar mellan modeller. Eftersom återanvändning av en modell redogörs i en child-modells tillhörande artikel, refereras då även artikeln som tillhör dess parent-modell. Genom att hämta de modellers artiklar som var fritt tillgängliga via PubMed Central [58], [59], kunde referenslistor och artikeltitlar programatiskt jämföras. Om matchande textstycken hittades, taggades de för vidare undersökning av en projektdeltagare.

För att underlätta projektdeltagarnas arbete skapades en funktion som för en given modell öppnade webbsidan till artikeln samt återgav alla potentiella kopplingar. Projektdeltagaren fick sedan utifrån artikeln utvärdera om kopplingen var reell och av vilken sort den var. I paradigm 1 utforskades också hur kopplingar skulle klassas, och detta lade grunden till den slutgiltiga klassificeringen i paradigm 2.

Med paradigm 1 av koden analyserades 143 artiklar tillhörande 179 modeller för kopplingar. Det framgick vid analysen att inga av modellerna inte var GEMs, så en bättre filtreringsmetod söktes. Efter handledning framgick att den mest optimala metoden var att filtrera sökningen efter begränsnings-baserade modeller. Emellertid innebar detta inte att endast GEMs ingick i filtreringen, vilket innebar att en extra kontroll av varje artikel implementerades. Detta sänkte mängden modelltillhörande artiklar tillgängliga via Biomodels till 87 och motiverade bytet till paradigm 2. Från citaten för kopplingarna i paradigm 1 skapades ordmoln för att se vilka ord som oftast associerades med en given koppling. Dessa ordmoln hittas i 3.2.

I paradigm 2 genererades inga potentiella kopplingar, utan varje artikel söktes igenom manuellt av projektdeltagarna. De genererade ordmolnen från paradigm 1 assisterade i sökandet. Ett gränssnitt för att annotera datasetet direkt i MATLAB skapades, samt en funktion för att konvertera MATLAB-formatet till JSON. För att kunna annotera datasetet samtidigt var konvertering essentiellt eftersom versionshanteringsverktyg kunde användas.

2.3 Manuellt adderade modeller till dataset

Utöver dessa 87 artiklar adderades även flera manuellt för att på så sätt stärka visualiseringen av GEMs. För att effektivt hitta fler kopplingar med tanke på tidskravet utgick sökningen från väl citerade modeller, såsom Recon 1. De artiklar som är åtkomliga på PubMed Central har en tillhörande lista av samtliga refererade artiklar, där den senast refererade artikeln ligger först. Utifrån denna lista identifierades de artiklar som uppenbart handlade om GEMs, alltså då titeln innehöll viktiga nyckelord påvisande detta.

Artiklarna analyserades med samma manuella sökmetodik som i paradigm 2. Då kopplingar hittades till den väl citerade modellen, alltså parent-modellen, adderades all relevant information av child-modellen till datasetet. Relevant information innefattar alltså fakta som rör modellen som lades till och även dess kopplingar till samtliga parent-modeller. Totalt adderades 20 modeller.

2.4 Visualisering av datasetet

För en initial överblick av datasetets GEM-släktträd gjordes först en visualisering i MATLAB. Där återgavs även histogram över ”modellpublicering per år” och ”modellåteranvändning per år”, som visas i avsnitt 3.3. Eftersom modellförbindelserna inte kunde åskådliggöras på ett tydligt sätt i MATLAB, utforskades visualiseringen vidare i programmen Cytoscape och Lucidchart [60], [61]. De tre olika verktygen hade varierande för- och nackdelar, men Lucidchart var mest implementerbar inom projektets tidsramar och valdes därför. Visualiseringen kan ses i Figur 3.1 där modellernas kopplingar, publiceringsår, modellnamn och grupperingar visas. Valda grupperingar är Fungi, Animalia och Bacteria, som redogörs för i avsnitt 3.1.

2.5 Planering, genomförande och informationshantering av intervjuer

Intervjuer kan användas i forskningssyfte för att komma fram till ny information. De är speciellt användbara när det finns ett behov av kvalitativ istället för kvantitativ datainsamling, särskilt när ämnet är komplext. Semi-strukturerade intervjuer används oftast i forskningssyfte för att kunna ställa specifika frågor till respondenter, samtidigt som upplägget kan ändras och nya frågor kan utvecklas under själva intervjun. [62] Av dessa anledningar har semi-strukturerade intervjuer valts som lämplig datainsamlingsmetod i detta projekt.

Fem forskare som direkt eller indirekt arbetar med GEMs valdes som intervjudeltagare, genom att använda den informations-baserade urvalsmetoden skapad för intervjuer [63]. Utifrån handledarnas rekommendationer valdes lämpliga intervjukandidater. För att komma fram till nya potentiella idéer och insikter valdes personer inom olika forskargrupper och med olika huvudforskningsområden. Det intervjuades tre doktorander, en postdoktor och en docent. Fyra av dem är anställda på Chalmers Tekniska Högskola (Chalmers), och den femte är anställd på Kungliga Tekniska Högskolan (KTH). Vissa av intervjudeltagarna arbetar direkt med att skapa GEMs, medan andra endast använder GEMs i sin forskning.

För att få detaljerade, individuella svar på specifika frågor valdes att intervjua en person i taget, istället för att bjuda in till gruppintervju. Intervjumetoden används för att underlätta både själva intervjuprocessen och datahanteringen efteråt, då det är enklare att urskilja vilken information som gavs av vilken respondent. [62] Det bestämdes att minst två projektdeltagare krävdes för att kunna utföra en produktiv intervju, eftersom sannolikheten att missa viktig information då skulle minska. Av

samma anledning bestämdes att spela in intervjuerna. För att skydda forskarnas integritet valdes det att inte spela in någon videofil, utan endast en ljudfil. Intervjuerna skedde via möten genom en videokommunikationsapplikation på nätet. Detta reducerade inte bara kostnader, utan underlättade också tidsplaneringen kraftigt. Intervjufrågorna och själva intervjuerna hölls på engelska.

När intervjufrågorna skapades, lades det mycket fokus på att komma fram till ny information som är relevant för projektet, och samtidigt få en bakgrund på individerna och deras användning av GEMs. Frågor som ”Tycker du att det här projektet är värdefull för GEM-forskningen?” och ”Har ditt arbete påverkats av att modellåteranvändningen inte finns åtkomlig på ett lättillgängligt sätt?” skapades för att direkt kunna påverka projektmålets utformning. Svaren utvärderades och jämfördes med befintlig information. Intervjufrågorna ligger i sin helhet i appendix A.1.1.

Sammanfattningar skrevs efter varje intervju, sådant att all information kunde delas med gruppen på ett effektivt sätt. Den inspelade ljudfilen var till stor hjälp i detta skede. Efter intervjuerna skapades också ett Excel-ark där all relevant information samlades. Detta gjordes för att kunna jämföra intervjudeltagarnas svar och åsikter på ett mer produktivt sätt. Informationen som kom fram under intervjuerna anges i tabellen 3.1.

3
Resultat och diskussion

3.1 Släktträdet över GEMs

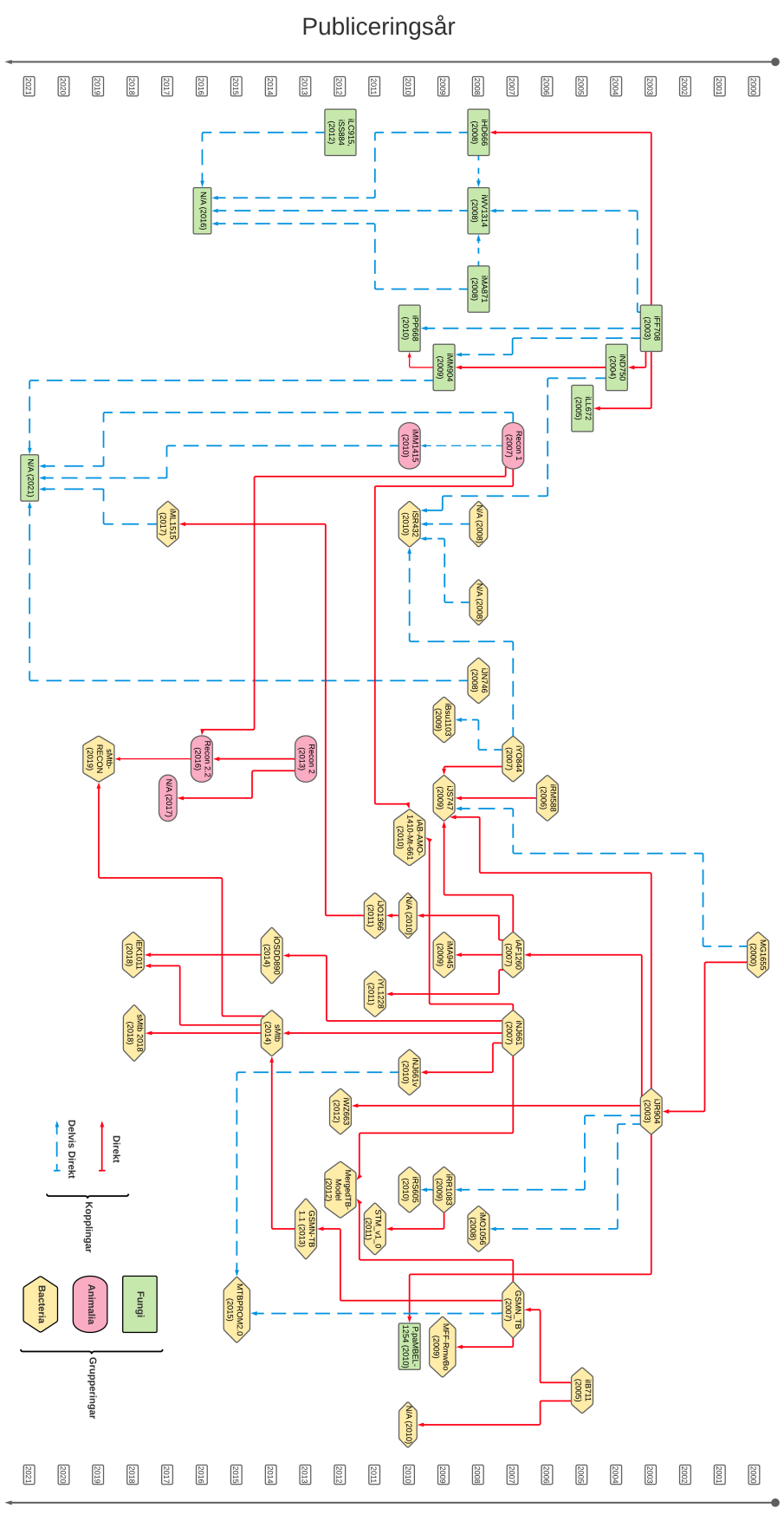
Projektets metodik, som redogörs för i avsnitt 2.4 resulterade i en visualisering av ett släktträd över GEMs, vilket kan ses i figur 3.1. Här påvisas de kopplingar, eller så kallade arv, som modellåteranvändningen utgör. Kopplingarna är antigen direkta eller delvis direkta, vilket klargörs i figuren. Kopplingarna skiljs åt genom användning av färgskillnad och linjetyp.

I figuren presenteras 52 modeller där deras modellnamn och deras kopplingar till varandra inkluderats. Modeller som inte har getts ett modellnamn av skaparna betecknas med N/A. Kopplingarna mellan modeller uppvisar direkta kopplingar med en tjock röd linje, medan delvis direkta kopplingar åskådliggöras med en streckad blå linje. Definitionen på kopplingarna anges i avsnitt 2.1. Modellerna är dessutom ordnade efter publiceringsår utifrån en axel som börjar på år 2000 och slutar på år 2021. Axeln är riktad nedåt i figuren.

Det finns tre olika nodtyper, som organiseras utifrån grupperingarna Animalia, Fungi och Bacteria. Då datasetet främst innehåller en stor variation av bakterier gjordes en gruppering av dessa. Vidare är det en majoritet av jästarter och modeller över människans metabolism, därav de resterande två grupperingarna. Några av de större bakteriefylum i trädet är aktinobakterier, firmicutes och proteobakterier. Av de 53 modellerna tillhör 36 modeller Bacteria, 12 tillhör Fungi och 5 tillhör Animalia.

I figuren kan ses att grupperingen Bacteria har flest antal direkta kopplingar, som exempelvis modell iNJ661 (2007) och iAF1260 (2007). En potentiell orsak skulle kunna vara att bakterier används som en ekonomisk gynnsam forskningsmodell. Vissa bakterier såsom *E. coli* är välstuderade och välkarakteriserade organismer [64]. Detta i sin tur skulle kunna vara en anledning till att GEM-forskning med bakterier är allmänt lättare, ty det funnits mer information och erfarenheter om dessa sedan 1990-talet, då forskning om GEMs startades [4].

Figuren visar dessutom att det finns kopplingar mellan parent- och child-modellernas olika grupperingar, det vill säga mellan organismer. Dessa kopplingar utgörs ofta av mekanism- eller reaktionsarv, vilket kan ses i figuren då det största antalet av kopplingarna är delvis direkta. En sådan modulär GEM-återanvändning verkar vara ett naturlig förlopp. I andra fall, såsom sMtb-RECON, handlar det om en patogen inuti en annan organism, vilket kräver en kombination av båda nätverken.



Figur 3.1: Släktträd över 52 Genome-Scale Metabolic Models (GEMs) med modellnamn och publiceringsår inkluderat. De röda, tjocka linjerna motsvarar en direkt koppling mellan modellerna, medan de blåa, streckade linjerna föreställer en delvis direkt koppling. Tre grupperingar har inkluderats: Animalia, Fungi och Bacteria.

3.2 Intervjuerna belyste viktiga faktorer av arbetet och viss problematik inom modellåteranvändning

Redan existerande GEMs uppdateras och utformas för att svara på nya frågor eller för att anpassa modellen beroende på applikationen. Om en tydlig återanvändning eller uppdatering av modellerna inte presenteras försvåras vidareutvecklingen och det fortsatta arbetet med GEMs. Detta problem är än idag aktuellt och påverkar forskare som genererar nya GEMs eller använder redan existerande sådana för eget syfte. Problematik kring att modeller ej kan jämföras på ett enkelt sätt och användas i simuleringsändamål diskuterades under intervjuerna.

Fem intervjuer genomfördes, många frågor ställdes och intensiva diskussioner efterföljdes. Relevant information samlades efter varje intervju och essentiella delar av den redovisas i tabell 3.1. Beslutet togs även att ej ange deltagarnas namn, istället benämns forskarna med bokstäverna A, B, C, D, E, vilka tilldelades slumpvist. Av de forskarna som deltog var konsensus av åsikterna att projektet är av värde, som doktorand B underströk; detta arbete har hittills varit ”a gap in the field”.

Forskare A, doktorand på Chalmers, ansåg att en visualisering av datasetet skulle kunna vara till stor hjälp, eftersom det i dagens läge ännu inte finns lättillgänglig information som berör modellåteranvändning. Forskare E, doktorand på KTH, påstod att en sådan visualisering skulle kunna bidra till att öka trovärdigheten hos GEMs. Det skulle förenkla processen att fastställa upphovet till de modeller som utvecklats hittills. Det kan även förverkliga en standardisering av GEMs i framtiden.

3.2.1 Påverkande faktorer vid en kvalitativ datainsamling

Semistrukturerade intervjuer planerades och genomfördes vilket klargörs i avsnitt 2.5. För att garantera kvalitet av ställdes strukturerade och öppna frågor, utifrån forskarnas föregående arbete eller artiklar som de deltog med. På sådant vis kunde forskarnas svar öppna nya spår för följdfrågor och därmed öka den relevanta informationen i diskussionen.

Med strukturerade frågor blir jämförelsen av data mer objektiv, och trovärdigheten av arbetet ökar. För att vidare minska risker för misstolkning eller missförstånd under intervjun skickades frågorna till deltagare i förväg. De kunde då förbereda sig inför intervjun, vilket ledde till mer utförliga och detaljerade svar.

Intervjuerna genomfördes individuellt med varje deltagare, trots att fyra av dem anställda jobbar på Chalmers och hade kunnat medverka i en gruppintervju. Forskarna gav inte bara insikt till relevanta parametrar, men också tankar och åsikter angående problem inom branschen. När relevanta parametrar lyfts fram diskuterades dessa mer utförligt, och klargjordes vilka som skulle inkluderas i projektet.

En gruppintervju skulle kunna höja kvaliteten av insamlad data, men kräver generellt mer tid då antalet deltagare ökas. I en sådan intervju skulle forskarna kunna

Tabell 3.1: En redovisning av frågor som ställdes under intervjuerna med dess svar.

	Forskare A	Forskare B	Forskare C	Forskare D	Forskare E
Bakgrund	PhD på CTH: Arbetar med att konstruera GEMs	Postdoc på CTH: Användare av GEMs för applikationer (t.ex. cancerforskning)	PhD på CTH: Arbetar med GEMs; adderar proteinsekretions-nätverk till GEMs	Docent på CTH: Projektledare Biology & Biological Engineering department	PhD på KTH: Användare av GEMs, arbetar med multi-omics-data
Vad jobbar forskarna med just nu?	Modellerar GEMs för flera olika arter (jäst, bakterier, mänsklig metabolism)	Använder mänskliga GEMs för att analysera multi-omics-data, hitta nya metoder att behandla cancer.	Jobbar med proteinsekretionsvägar.	Genererar nya GEMs och kurerar befintliga GEMs till nya organismer eller utveckla befintliga modeller.	Analyserar transkriptomiska data för hjärt-kärlsjukdom med användning av data från flera mänskliga vävnader.
Anledning till att göra intervjun	Publicerat en uppsats som berör samma ämnesområde och problematik som detta arbete.	Tyckte att projektet är ”part of a bigger picture” och skulle kunna gynna fler forskare.	Dels för att dela med sig av erfarenheter, och dels för att informera om möjligheter av modellen som han jobbar på.	Intresserad av att hjälpa studenter framåt i deras projekt.	Ville bidra till ämnet och utöka området eftersom GEMs kommer vara viktigt för forskning inom Systembiologin.
Tankar angående projektet	Eftersom många modeller härrör från andra modeller skulle ett visualiserings-verktyg vara till stor hjälp.	Projektet är “a gap in the field” som skulle kunna besvara frågan "which GEMs are best for what?" som kan ge mycket värde.	Det är uppenbart för de forskare som jobbar inom branschen att modellåteranvändning är viktigt och att projektet är värdefullt.	Påpekade att exempelvis ett för generellt flödesschema som visualisering ej ger tillräcklig information. Ex. metabolit-ID behöver inkluderas för ökat värde.	Projektet kan potentiellt hjälpa till att öka trovärdigheten hos GEMs, eftersom det blir enklare att granska härkomst. "What GEM is best for what?". Även standardisering av GEMs kan vara möjlig i framtiden.
Är modellåteranvändning relevant för dem?	Ja: Mycket relevant, då information om modellåteranvändning är inte lättillgänglig idag	Inte speciellt, eftersom använder de mest aktuella Chalmers-utvecklade GEMs som standard.	Ja: Det är en stor utmaning att skapa modeller och hitta relevant information. Äldre modeller kan användas men kvalitetsinformation är inte alltid tillgänglig.	Det beror på vilken information som ingår i visualiseringsverktyget, ett flödesschema har begränsande värde.	Ja: Det är svårt att spåra GEMs, och svårt att samla in information om vilken modell som är bäst. Modellåteranvändning skulle kunna hjälpa till med standardisering av GEMs.
Övriga viktiga annoterings-parametrar?	·Versionshantering ·Metabolit-ID ·Interoperabilitet (dataintegration)	·Versionshantering ·Gruppering av reaktioner och metaboliter i olika delsystem ·Kvalitetskontroll	·Jämförelse (reaktions-ID, metabolit-ID) ·Jämförelse mellan olika modeller	· Versionshantering · Metabolit-ID ·Interoperabilitet ·Jämförelse mellan olika modeller	· Metabolit-ID och andra ID. Även försöka hitta konsensus mellan dessa. ·Interoperabilitet
Andra aspekter att fokusera på?			·Signaleringsvägar Dessa är inte en del av GEMs än, men det skulle kunna vara en intressant aspekt.	· Kvalitetskontroll. (Om en äldre modell återanvänds oftare, är de bättre än nya modeller?) ·Eftersom ingen konsensus finns ang. ID skapas problem, men behöver hitta en balans mellan ett för generellt flödesschema och ett för djupgående verktyg med metabolit-ID. Bör anpassas till projektets mål.	

stödja eller bestrida varandras åsikter, vilket nyanserar diskussionen. Metoden rekommenderas i kombination med individuella intervjuer för framtida projekt för ett mer omfattande resultat.

Dessutom kan kvaliteten på intervjuresultatet höjas genom att dels öka antal intervjudeltagare, och dels intervjua forskare med olika akademiska bakgrunder inom det relevanta området. Att öka antalet intervjudeltagare hjälper till att förstå vilken information som är av största intresse bland forskare. De flesta intervjudeltagarna är anställda på Chalmers men använder sig av GEMs på olika sätt. Följden blir att värdet av den insamlade datan kan ifrågasättas då forskarnas perspektiv blir mycket lika. Emellertid hade intervjudeltagarna olika anställningar och expertis såsom doktorand och docent, vilket kan argumenteras lyfter kvaliteten på intervjuerna då perspektiven på så sätt skiljer sig. För att uppnå maximal värde av intervjuresul-

tatet rekommenderas det att intervjua forskare från olika institutioner med olika tillämpningsområden av GEMs.

3.2.2 Forskarnas åsikter kring annoteringen ”modellåteranvändning” och aktuella annoteringsparametrar

Trots att intervjudeltagarna arbetar inom olika forskningsområde och har olika akademisk bakgrund nämnde alla de otydliga och oorganiserade kopplingarna mellan GEMs som en olägenhet. Fyra av fem deltagare ansåg att ”modellåteranvändning” är en relevant annotering att låta projektet utgå ifrån. Forskare D, som är docent på Chalmers, framhävde dock att annoteringen endast är av betydelse om det medföljer mer djupgående information gällande modellerna.

Respondent D menar att ett flödesschema som endast visar var modellerna härstammar ifrån inte är till någon större hjälp för forskare inom branschen. Istället bör mer detaljerad information ingå för en ökad förståelse om hur modellerna återanvänds och vilka sorts kopplingar detta utgör. Bland annat bör informationen visa på i vilken omfattning GEMs används, specifikt när det rör modellarv. Med andra ord, om den fullständiga parent-modellen använts som bas för den nya child-modellen borde detta tydligt klargöras, och tvärtom bör det också redogöras om endast vissa modellkomponenter inkluderats.

En central aspekt som diskuterades i varje intervju var valet av relevanta annoteringsparametrar och vilka som ansågs vara essentiella för projektet som helhet, men också för branschen som forskarna arbetar inom. De parametrar som diskuterats under projektets gång och som även diskuterades under intervjuerna är modellarv, modellversioner (versionshantering) och ID av olika modellkomponenter.

3.2.2.1 Relevans och implementerbarhet av parametrarna ”modellarv” och ”modellversion”

Parametern ”modellarv” ansågs av de flesta intervjudeltagarna tillföra ett intressant perspektiv till modellåteranvändningen som den ser ut idag. Däremot fanns det en viss osäkerhet i vad exakt modellarv betyder, speciellt var det forskare D som var intresserad av en utförligare förklaring av benämningen. Projektdeltagarna som deltog under just den intervjun klargjorde att parametern i fråga är kopplad till projektets huvudmål: en visualisering av ett släktträd över GEMs. Kopplingar mellan äldre och nyare modeller kan alltså ses som ett arv. Även här betonades av respondent D att mer information än bara modellarv krävs för att släktträdet ska vara av betydelse. Detta kan bero på att forskare D, som är docent på Chalmers, innesitter mer expertis och erfarenhet inom området än de andra respondenterna, vilka är doktorander och postdok.

Sammantaget ansågs det emellertid att parametern modellarv är högst relevant för projektet och för de olika forskarnas områden. Likvisst diskuterades det i samtliga intervjuer om att annoteringsparametern ”modellversion” kan innebära ett ökat värde av projektet.

Intervjudeltagarnas åsikter belyste på samma sätt som Lieven m.fl. [51] att versionshantering av modeller är en mer testdriven metodik för hur GEMs konstrueras och tillämpas. I enlighet med avsnitt 1.1.1.5 ansåg forskarna att modellarv är skiljaktigt från modellversioner och hanteringen av dessa. I synnerhet när det gäller modellarv publiceras endast en slutgiltig modell med tillhörande vetenskaplig artikel. Följaktligen finns det inte någon mer djupgående information tillgänglig att granska, vilket å andra sidan versionshantering möjliggör. Just den här mer enhetliga information kan nyttjas till fördel, som forskarna under intervjuerna belyste.

Versionshantering tillåter att ändringar som vidtas i modellversioner kan spåras och granskas av forskare som aktuellt arbetar med modellerna, men även av utomstående parter. Utifrån tillgängliga versioner kan alltså information erhållas angående vem som utför ändringar och vid vilken tidpunkt. Detta möjliggör för forskarna att analysera i vilken tillämpning en specifik version har använts och hur väl den kan implementeras i liknande eller andra skilda syften.

Tillgängligheten av versioner ger alltså forskare tillfället att utvärdera om den senaste versionen av en GEM skall vidareutvecklas eller om tidigare versioner fungerat bättre. Detta skulle kunna ge en indikation på kvalitet, vilket också överstämmer med kommentarerna från forskare C och D, som visas i tabell 3.1. GitHub, som också används i detta projekt, är ett exempel på ett verktyg där versionshantering kan behandlas på ett strukturerat och öppet sätt. Genom att nyttja detta eller liknande verktyg erhålls en ändrings- och utvecklingshistorik av GEMs.

Som det däremot klargörs för i avsnitt 1.5.2 som rör annoteringsparametrar, togs ett beslut att exkludera modellversion som parameter och istället lägga fokus på modellarv. Då versionshantering ej implementeras i samma omfattning på alla institutioner eller av alla individer som konstruerar GEMs, skulle datainsamlingen och analys av modellerna bli alltför komplex och tidskrävande för detta projekt.

Emellertid bör det understrykas att modellversioner är högst relevant för modellarv, då det bland annat ger möjligheten att på ett enklare sätt jämföra modellkomponenter och tillämplighet av en specifik version av en GEM. Sammantaget skulle detta i framtiden alltså kunna inkluderas i datasetet för att öka dess värde och funktionalitet. I själva verket skulle en mer informationsrik visualisering av datasetet, där modellversioner är inkluderade, kunna visa fördelarna med just versionshantering. På så sätt skulle detta få andra forskare och institutioner som idag ej nyttjar hantering och åtkomst av modellversioner, att börja implementera denna metodik.

3.2.2.2 ID för olika modellkomponenter innebär ett ökat värde av visualiseringen

I intervjuerna framgick tydligt en central aspekt gällande de olika komponenterna en modell är uppbyggd av, alltså metaboliter, reaktioner, cellkompartement och gener. Som mer ingående förklaras i avsnitt 1.1.1.3 benämns dessa med hjälp av olika identifierare, såsom metabolit-ID, reaktions-ID och så vidare. Samtliga intervjudeltagare betonade att ID är både en högst essentiell del av rekonstruktion och återanvändning av GEMs, men också en problematisk aspekt inom området. Forskarnas åsikter

stämde nämligen överens med studier [45], [44] som pekar på en stor heterogenitet för en given komponents identifierare och svårigheterna detta skapar (se avsnitt 1.1.1). Således ansågs ID som både en möjlig och värdefull annoteringsparameter, trots vissa faktorer som försvårar detta.

Främst diskuterades metabolit-ID som en mycket viktig annoteringsparameter. Då modeller antingen konstrueras manuellt eller med hjälp av ett modelleringsvertyg skapas en stor variation inom identifierare av metaboliter. Detta oavsett om det gäller grundläggande modelldelar, såsom ATP och DNA, eller mer unika för just metabolismen i den organism modellen representerar. På samma gång som samtliga intervjudeltagare belyste vikten av metabolit-ID, underströk främst forskare D att projektet ej bör syfta till att ”solve other peoples problems”. Här spelar nämligen problematiken kring standardisering in, som till viss del förklaras i avsnitt 1.1.1.4, vilket skulle ha gjort projektet alltför komplext och tidskrävande. Speciellt i avseende på vart informationen om ID finns tillgänglig, generellt i SBML-filerna, och att variationen av ID är såpass stor.

Från projektets start utvecklades metodiken att samla och analysera data utifrån modellers tillhörande artiklar, vilket därför innebar att en anpassning av både det manuella och automatiserade arbetet inte var möjlig. Däremot diskuterades som redan nämnts specifikt metabolit-ID som en värdefull aspekt till modellarv, då det är en essentiell faktor när det rör konstruering av GEMs. Genom att inkludera ID skulle det bli möjligt att systematiskt kombinera, jämföra och utvärdera modeller på ett enklare och mer effektivt sätt, vilket skulle gynna branschen. Eftersom det är nära ihopkopplat med standardisering av GEMs (se avsnitt 1.1.1.4), skulle även detta gynnas av mer transparens och lättillgänglighet inom identifierare.

3.2.3 Andra förslag som eventuellt är implementerbara i framtida projekt

Intervjudeltagarna föreslog också relevanta aspekter som ofta dyker upp som ett hinder inom branschen, såsom interoperabilitet och kvalitetskontroll. Den heltäckande analysen som skulle krävas för att inkludera aspekterna i datasetet, är av en mer avancerad karaktär än exempelvis för modellåteranvändning. Det kräver också djup förståelse av modellerna vilket är tidskrävande och erfordrar en mycket större mängd datainsamling. Med hänsyn till detta, kan förslagen eventuellt implementeras i framtida projekt med bredare eller utförligare omfattning.

3.2.3.1 Interoperabilitet och databasintegration

Interoperabilitet är ett mycket komplext ämne med många olika aspekter, bland annat databasintegration. Enligt forskare E ska integration till externa databaser vara mer omfattande i ändamål att kombinera multiomics-data med modellerna, något som försvåras då ingen konsensus finns vid namngivning av ID. När identifierare för en komponent skiljer sig i en och samma modell kommer de betraktas som separata komponenter i beräkningar, vilket också sänker kvalitetskontrollen. På sådant vis ligger en stor vikt på att standardisera ID av modellkomponenter innan alla hittills

konstruerade GEMs kan integreras till databaser.

Användningen av ett standardfilformat, tidigare nämnt i avsnitt 1.1.1.4, kan också vara en lösning på problematiken som uppstår med interoperabilitet. Genom att ersätta alla dessa unika sammanhörande filformat och identifierare med unika standardiserade format för båda faktorer kan i sin tur de fullständiga modellerna standardiseras. Detta möjliggör resultatrik databasintegration och höjer modellens kvalitet.

3.2.3.2 Kvalitetskontroll

En annan annotering som föreslogs av vissa forskare var kvalitetskontroll. Ett exempel på dess användning som framlyftes ett flertal gånger var jämförelse mellan olika modeller. En sådan jämförelse kan genomföras över flera aspekter, till exempel hur ofta modellerna används, eller om det finns vissa fel i en parent-modell som en child-modell är lösning på. Dessutom lämpar sig olika child-modeller från samma parent-modell olika bra för en given tillämpning. Forskare B nämnde också att det är svårt att välja mellan olika redan konstruerade GEMs, speciellt om de härstammar från samma parent-modell, då sättet de skiljer sig på inte framgår tydligt.

Den detaljerade analyseringen av befintliga modellerna för att förstå skillnaderna mellan dem kan underlätta användningen av GEMs. En lösning på kvalitetskontroll som kan kvalitettesta modeller är tillämningsprogrammet MEMOTE, som också presenteras i Lieven m.fl.:s arbete. [51] Programvaran erbjuder flera kvalitetskontroller. Bland annat annoteringstester för att säkerställa att annoteringen sker enligt gemenskapsstandarder och andra tester såsom närvaroverifiering av komponenter och korrekthet av en modell. Detta tillåter jämförelse mellan givna modeller. Därtill kommer MEMOTE tilldela en poäng utifrån kvantifieringen och kondenseringen av testresultatet. På så sätt genomförs standardiseringen av modeller.

3.2.3.3 Inverkan av signaleringsvägar

Utöver annoteringarna nämnde forskare C signaleringsvägar som påverkande faktor i skapandet av GEMs. Forskaren beskrev hur signaleringsvägar är kemiska seriereaktioner i cellen, och förklarade även att implementering av signaleringsvägar ej nämns då en GEM genereras, vilket kan sänka forskningsarbetets kvalitet. Tillämpningen av modeller kan då bero på om det finns en specifik pågående signaleringsväg i cellen eller inte. Forskare C menar att tillägget av denna information till modeller är viktig för utökning och utveckling av GEMs.

3.2.4 Vidareutvecklingen av GEMs i framtiden

Genom att lättillgängliggöra beskrivningen av modellens återanvändning kan forskares arbete och vidare utveckling av GEMs underlättas, i enlighet med FAIR-principerna. Trots en övervägande positiv respons gällande valet av annoteringen ”modellåteranvändning”, fanns det dock skilda åsikter. Problematiken som uppstått av de många modelleringsverktygen och databaserna, den stora heterogeniteten inom ID, olika val av filformat (såsom SBML), olika metod gällande versionshantering

och alla andra påverkande faktorer, är idag svårlöst. Emellertid utgör den nuvarande situationen, där återanvändningen av GEMs som till viss del är otillgänglig och många gånger svårnavigerad, ett onödigt komplicerat område inom forskningen. Det finns otaliga sätt att angripa de svårigheter och problem som idag är aktuella och i behov av att behandlas. Utifrån intervjuerna belystes dock de mest angelägna frågorna.

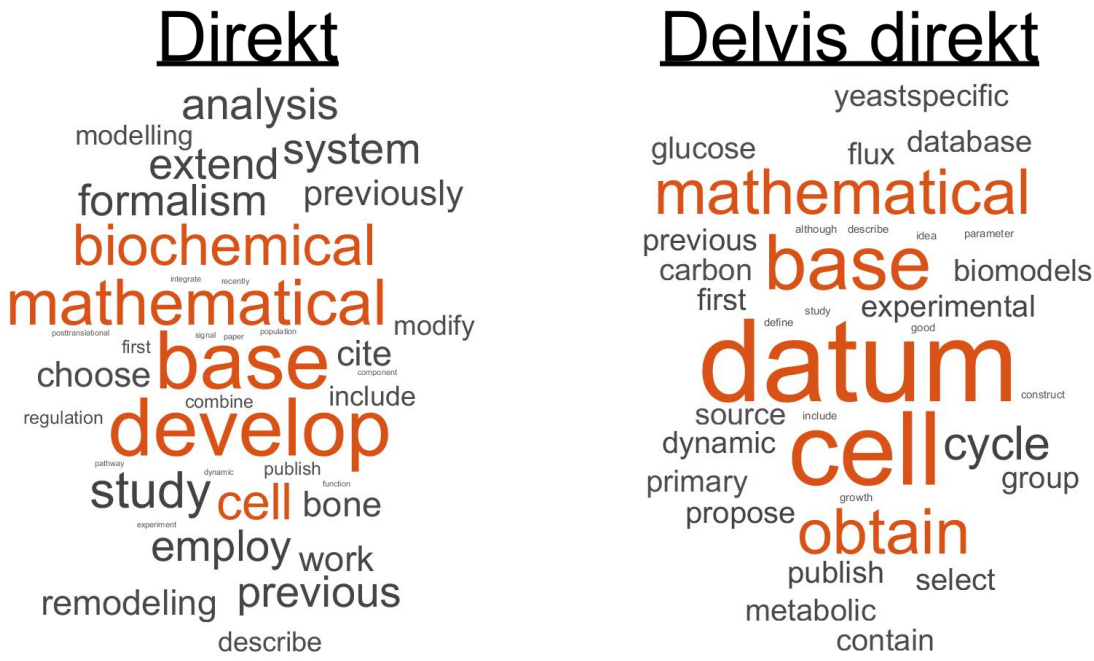
Precis som studier påpekar kan modellåteranvändning gynna och förenkla forskares egna arbete med GEMs, men också lägga grunden för ett mer öppet samtal och samarbete. Övervägande åsikter kring modellarv var positiva, men som redan nämnts kan en mer informationsrik visualisering ha ökat värde. För att öka värdet och användbarheten av släktträdet skulle därför modellversioner och hanteringen av dessa, ID av olika modellkomponenter och filformat kunna inkluderas som parametrar. En mer informerad modellåteranvändning kan också gynna standardiseringen av GEMs, som i sin tur kan öka trovärdigheten och kvaliteten av modeller.

3.3 Annoteringsmetodik i paradigm 1 och 2

Projektets annotering i paradigm 1 inkluderade inga GEMs då filtreringsmetoderna uteslöt dem, som nämnts i avsnitt 2.2. När de nya filtreringsmetoderna implementerades i paradigm 2 sänktes mängden modelltillhörande artiklar tillgängliga via Biomodels till 87. Vid detta stadie i projektet ansågs det finnas två metoder att välja mellan.

Det första alternativet var att använda samma metod som paradigm 1, med automatisk eller manuell nedladdning av artiklar. Programatisk nedladdning av artiklar har begränsad tillgänglighet, och det nya datasetet var en tiondel av storleken i paradigm 1. Med detta i åtanke uppkom en ny metodik som uteslöt koden för att söka i artiklar. Istället grundade den sig i manuell sökning av artiklarna, eftersom en manuell sökning antas vara av högre kvalitet än en automatiserad. Ett antal hjälpmedel och metoder utformades som följd av paradigmskiftet. En av dessa var ordmoln av de främst förekommande orden i ”direkt” och ”delvis direkt” kopplingarna som skapades från de 143 bearbetade artiklarna från det första stadiet i projektet. Dessa assisterade i analysen av artiklarna, och ses i figur 3.2.

Figur 3.2: Ordmoln från paradigm 1



Problemen som paradigm 2 medförde, huvudsakligen arbetsbörda, var inte av stor påverkan med ett litet dataset som vid denna tidspunkt innehöll 87 artiklar tillhörande 162 modeller. Därför valdes den som den slutliga metoden. För bearbetning av ett större dataset än omfattningen av det här projektet rekommenderas en programatisk metod, förslagsvis baserad på maskininlärning. Projektets dataset kan då vara startpunkten för programmets upplärning. Koden som använts i paradigm 1 skulle också kunna vidareutvecklas i ett framtida projekt.

För att modellernas kopplingar ska vara pålitliga måste metodiken för granskningen vara transparent. Ett sätt är genom användning av de citaten som medföljer kopplingsparametern i det här projektet. Utifrån citaten kan kopplingens sort klargöras eller omvärderas i enlighet med FAIR-principerna.

3.4 Diskussion kring frågor

I denna sektion diskuteras frågorna som togs upp i avsnitt 1.3. Projektets ursprungliga mål var att besvara dessa frågor utifrån det annoterade datasetet och visualiseringen, men under arbetets gång förändrades de frågor som var av intresse eller som det fanns möjlighet att svara på. Delvis på grund av begränsningar i metoderna, och delvis med insikt från intervjuerna. Nedan följer en diskussion om förändringarna och svaren till de nya frågorna utifrån alla aspekter av projektet.

Hur stor andel av modellerna har någon koppling?

På grund av projektets mindre omfattning krävdes följande omformulering av frågan: Hur många av de 107 artiklar som inkluderades i datasetet hade en eller flera kopplingar? Svaret är att 73 utav de 107 artiklarna i datasetet hade en eller flera kopplingar. Eftersom datasetet till stor del endast plockats från BioModels kan resultatet inte antas representera en trend för artiklar från andra databaser.

Finns det modeller som återanvänds mer än andra, och indikerar det en högre kvalitet?

Utifrån visualiseringen i figur 3.1 är det tydligt att det finns modeller som återanvänds mer än andra. Men det är mer relevant att fråga: Vilka modeller i datasetet återanvänds mer än andra? Ett exempel på detta är modellerna iFF708(2003) och iJR904(2003), där båda har sex kopplingar till sina child-modeller. iFF708 har tre stycken direkta och tre stycken delvis direkta kopplingar, medan iJR904 har fyra direkta respektive två stycken delvis direkta kopplingar. De äldre modellerna återvänds oftare än de nya vilket kan bero på att de har haft längre tid att inkorporeras i nya modeller, eller att GEMs konkurrerar med varandra för kopplingar allt eftersom deras antal ökar. Då det inte gjorts någon undersökning av modellkvalitet så kan det inte indikeras hur återanvändning påverkar denna aspekt av modellerna.

En följdfråga till frågan ovan är: Vad säger mängden kopplingar som är direkta jämfört med delvis direkta på kvaliteten av child- eller parent-modellen? Utifrån hur direkt och delvis direkta kopplingar har definierats i detta projekt medförs ingen högre kvalitet med helt direkta kopplingar. Exempel på detta är en delvis direkt koppling där en del reaktioner och metaboliter tagits för att rekonstruera en ny modell för en organism med större genetisk skillnad än den ursprungliga organism, vilket inte skulle indikera låg kvalitet i parent-modellen.

Hur ofta uppdateras modeller?

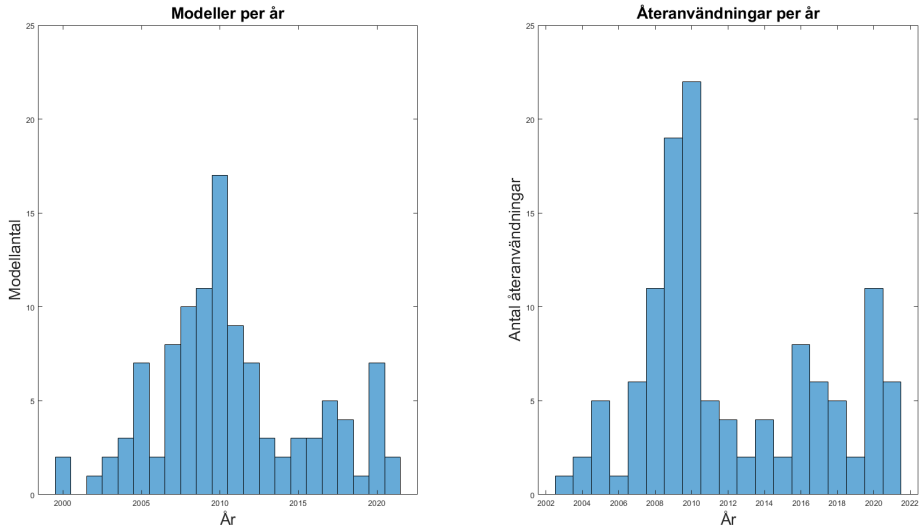
Som beskrivs i avsnitt 1.1.1.5 och 3.2.2.1 är modelluppdatering främst applicerbart för versionshanterade GEMs. Eftersom metoden i det här projektet utgår från modellernas artiklar fanns ingen möjlighet att beskriva modelluppdatering och parametern uteslöts. I 4.1 diskuteras möjligheter för att inkludera den i framtida projekt.

Hur ser trenden ut för antalet publicerade modeller per år?

Antalet publicerade modeller per år i datasetet redovisas i figur 3.3 nedan. En ökande trend i publicering kan observeras mellan 2000-2010. Därefter har trenden varierat mellan 2010-2020, men om de här trenderna kan extrapoleras till andra databaser är inte säkert.

Har det skett en förändring i antalet modeller som återanvänds per rekonstruerad modell genom åren?

Utifrån trenderna som observerats i datasetet (se figur 3.3) följer återanvändningen i stort sett publiceringen. Återanvändningstrenden har dock en större varians, då datasetets storlek inte är tillräcklig för att kompensera för variansen i de individuella modellerna. Som tidigare kan dessa trender ej extrapoleras till en större skala.



Figur 3.3: Här redovisas två histogram som plottades med MATLAB. Den västra figuren beskriver antalet publicerade modeller över de senaste 20 åren. Figuren till höger visar en trend på antalet återvändningar som modeller har över åren.

4

Slutsats

I det här projektet har det skapats ett annoterat dataset av 107 artiklar tillhörande 182 GEMs, och hälften av dem kunde kopplas till ett centralt släktträd. Med intervjuerna som underlag är förhoppningen att verksamma forskare inom systembiologi kommer kunna använda resultaten i den här rapporten för att underlätta användningen och konstruktionen av GEMs. Fyra av fem målfrågor har även besvarats, om än anpassade till projektets omfattning.

En stor del av projektet utgjorde att utvärdera och integrera åsikter och tankar från forskare som aktuellt arbetar med GEMs. Huvudsakligen eftersöktes om de ansåg arbetet och visualiseringen av modellåteranvändning vara värdefulla för branschen och sammantaget var åsikterna positiva. Kort sagt ansågs att ett projekt som detta saknats hitintills, och att det har betydelse för utvecklingen av GEMs. Annoteringen ”modellarv” och dess innebörd för släktträdet över GEMs utgör därför en påbörjad grund för ett fortsatt arbete där fler modeller kan adderas.

4.1 Förslag på förbättringar av projektet och framtida frågeställningar om GEMs

För att försäkra att åsikterna är representativa av hela systembiologi-fältet bör en större mängd forskare och andra GEM-användare ingå i intervjuerna. Med fler respondenter fångas även fler specialfall upp, såsom inverkan av signalvägar, se avsnitt 3.2.3.3. Ett framtida projekt kan även genomföra gruppintervjuer för att ge en mer nyanserad åsikt, men det skulle vara mer tidskrävande.

Utifrån diskussion med forskarna framkom många aktuella och viktiga faktorer som skulle kunna öka värdet och användbarheten av visualiseringen, såsom modellversioner och ID av modellkomponenter. Potentiellt kan åsikterna färgas av att fyra av fem av intervjudeltagare arbetar på Chalmers. Däremot belyser studier, se avsnitt 1.1.1, samma faktorer och problem som intervjudeltagarna förde på tal och av detta skäl bedömdes forskarnas åsikter vara tillämpbara.

Av dessa faktorer framlyftes modellversioner som mest essentiell, eftersom det i nuläget finns många GEMs som inte bara återanvänds i andra modellkonstruktioner, utan också har ett flertal versioner. Versionshantering av modeller anses vara betydelsefullt och viktigt eftersom arbetet blir mer testdrivet. Konstruktionen av modellversioner blir även transparent då en historik skapas av de förändringar som utförs. Således sammankopplas detta väl med FAIR-principerna som presenterades

som en central aspekt av projektets syfte. Modellversioner redovisas inte likadant för versionshanterade modeller och publicerade modeller, så en kompromiss skulle behöva göras för att inkludera båda i ett framtida projekt.

Fokus borde även läggas på ID av modellkomponenter som forskarna lyfte fram, men även filformat som vissa studier diskuterade, se avsnitt 1.1.1. Genom att inkorporera dessa två aspekter ökar informationsinnehållet av modeller, vilket i sin tur kan gynna standardiseringen av GEMs. Med standardiserade modeller kan kvalitetskontroll, exempelvis genom användning av MEMOTE, genomföras och härav följer mer trovärdighet och funktionalitet av GEMs. Problematiken kring heterogenitet av ID och att SBML-filformat ej är standard kvarstår dock. Sammantaget kan emellertid ett öppet och informativt samtal angående dessa faktorer främja modellåteranvändning där ID och filformat tillämpas.

En framtida frågeställning om GEMs som belystes under intervjuerna är hur signaleringsvägar kan inkorporeras till metaboliska modeller, (se avsnitt 3.2.3.3). Just detta skulle alltså snarare vara en förbättring av modellerna som funktionella verktyg inom forskningen och inte något som bör inkorporeras i det här eller liknande projekt.

För att kunna extrapolera trender av datasetet till hela fältet krävs att fler modeller behandlas. Förutom skillnader i metodik är det även relevant att hitta modeller att analysera. Då konstruktionen av GEMs kan variera, bör komposition av datasetet vara annorlunda för ett givet projekt. Om intresset är att bygga upp ett släktträd och undersöka hur modeller återanvänds kommer modeller som är relativt mer manuellt konstruerade vara relevanta. Analyser av identifierare och kvalitetskontroll kommer istället fokusera mer på automatiskt och semi-automatiskt genererade modeller, då de utgör en majoritet. När ett släktträd byggs upp blir ett större dataset en stor fördel då risken att missa modeller som länkar grenar av trädet minskar.

Släktträdet av GEMs i det här projektet är grupperat efter Animalia, Fungi och Bacteria, men framtida projekt bör utveckla visualiseringen och grupperingarna på sådant vis att arter skiljs åt medan deras genetiska närhet synliggörs. I det här projektet antogs även att varje modell är av en enskild organism eller dess delar, men det finns även GEMs som modellerar interaktionen mellan organismer såsom patogen och människa. Framtida projekt bör reflektera över hur en sådan modell ska passas in i taxonindelningen av datasetet.

Metod och koden som användes i det här projektet kan användas i eller bli inspiration för framtida projekt. I dagsläget lämpar sig inte metoden för större dataset, främst på grund av mängden manuellt arbete och databaserna som används. De manuella metoderna kräver inga program, och är därför simpla, men tidskrävande. Genom att automatisera kringliggande arbete kan dock tidsbelastningen minska. Detta kan till exempel involvera utvecklingen av gränssnitt. Om potentiella kopplingar genereras som i paradigm 1 skulle en mer stabil och större automation av de potentiella kopplingarna, så att användaren endast behöver klassificera de givna citaten, minska mängden manuellt arbete. Efter en andel modeller har bearbetats skulle ett maskinlärnings-program kunna utgå från dem för att klassificera resten

av kopplingarna. Det här projektets dataset kan då användas som utgångspunkt.

För att expandera datasetet behöver programmet integreras till en mer omfattande databas av artiklar, såsom SCOPUS. Då krävs att modellerna kan kopplas till sina respektive artiklar, förslagsvis genom DOI. Sedan behövs även tillgång till en licens som tillåter datautvinning på den givna databasen samt implementation av licensen i koden.

Avslutningsvis kan det konstateras att projektet är både värdefullt och innebär en påbörjad grund till en fortsatt utveckling av att koppla modeller utifrån modellåteranvändningen och det modellarv detta utgör. Det finns ett flertal förbättringar av varierande komplexitet och krav på insamling och bearbetning av data. Flera faktorer av de förbättringar som sammanfattats ovan kräver också att områden inom systembiologi-branschen utvecklas och standardiseras. Emellertid kan projekt som detta, i enlighet med FAIR-principerna, gynna ett mer transparent och öppet samtal inom detta forskningsområde och på så sätt främja en fortsatt förbättring och utveckling av GEMs.

Litteratur

[1] A. Emwas, R. Salek, J. Griffin och J. Merzaban, "NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations", *Metabolomics*, årg. 9, s. 1048–1072, 2013. DOI: 10.1007/s11306-013-0524-y.

[2] R. DeBerardinis och C. Thompson, "Cellular metabolism and disease: what do metabolic outliers teach us?", *Cell*, årg. 148, nr 6, s. 1032–1144, 2012. DOI: 10.1016/j.cell.2012.02.032.

[3] U. Sauer, "Metabolic networks in motion: 13C-based flux analysis", *Mol Syst Biol*, årg. 2, nr 62, 2006. DOI: 10.1038/msb4100109.

[4] C. Gu, G. B. Kim, W. J. Kim, H. U. Kim och S. Y. Lee, "Current status and applications of genome-scale metabolic models", *Genome Biology*, årg. 20, nr 121, 2019. DOI: 10.1186/s13059-019-1730-3.

[5] J. L. Robinson, P. Kocabaş, H. Wang, P.-E. Cholley, D. Cook m. fl., "An atlas of human metabolism", *Science signaling*, årg. 13, nr 624, 2020. DOI: 10.1126/scisignal.aaz1482.

[6] S. Mendoza, B. Olivier, D. Molenaar och B. Teusink, "A systematic assessment of current genome-scale metabolic reconstruction tools", *Genome Biol*, årg. 20, nr 158, 2019. DOI: 10.1186/s13059-019-1769-1.

[7] A. A. Labena, Y.-Z. Gao, C. Dong, H.-l. Hua och F.-B. Guo, "Metabolic pathway databases and model repositories", *Quantitative Biology*, årg. 6, nr 1, s. 30, 2018. DOI: 10.1007/s40484-017-0108-3.

[8] P. Karp och R. Caspi, "A survey of metabolic databases emphasizing the MetaCyc family", *Arch Toxicol*, årg. 85, nr 9, s. 1015–1033, 2011. DOI: 10.1007/s00204-011-0705-2.

[9] I. Schomburg, L. Jeske, M. Ulbrich, S. Placzek, A. Chang och D. Schomburg, "The BRENDA enzyme information system–From a database to an expert system", *Journal of Biotechnology*, årg. 261, s. 194–206, 2017, Bioinformatics Solutions for Big Data Analysis in Life Sciences presented by the German Network for Bioinformatics Infrastructure, ISSN: 0168-1656. DOI: https://doi.org/10.1016/j.jbiotec.2017.04.020.

[10] KEGG: Kyoto Encyclopedia of Genes and Genomes, *KEGG Release Notes*, https://www.genome.jp/kegg/docs/relnote.html, Online; 1 Mars 2021.

[11] C. Krieger, P. Zhang, L. Mueller, A. Wang, S. Paley m. fl., "MetaCyc: a multiorganism database of metabolic pathways and enzymes", *Nucleic Acids Res*, årg. 32, nr Database issue, s. 194–206, 2004. DOI: 10.1093/nar/gkh100.

[12] Rat Genome Database, *About Us*, https://rgd.mcw.edu/wg/about-us/, Online; 1 Mars 2021.

[13] C. J. Norsigian, N. Pusarla, J. L. McConn, J. T. Yurkovich, A. Dräger m. fl., "BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree", *Nucleic Acids Research*, årg. 48, nr D1, s. D402–D406, 2019. DOI: 10.1093/nar/gkz1054.

[14] P. D. Karp, "A survey of metabolic databases emphasizing the MetaCyc family", *Archives of toxicology*, årg. 85, nr 9, s. 1015–33, 2011. DOI: 10.1007/s00204-011-0705-2.

[15] R. S. Malik-Sheriff, M. Glont, T. V. N. Nguyen, K. Tiwari, M. G. Roberts m. fl., "BioModels—15 years of sharing computational models in life science", *Nucleic Acids Research*, årg. 48, nr D1, s. D407–D415, 2019. DOI: 10.1093/nar/gkz1055.

[16] J. Kim, J. Jun, Y. Kim, S. Chae, M. Roh m. fl., "BIOSILICO: A Biochemical Database Retrieval System", *Genome Informatics*, årg. 14, s. 693–694, 2003. DOI: 10.11234/gi1990.14.693.

[17] BsubCyc, *BSUBCYC UPDATE HISTORY*, https://bsubcyc.org/release-notes.shtml, Online; 1 MArs 2021.

[18] A. Van Moerkercke, M. Fabris, J. Pollier, G. Baart, S. Rombauts m. fl., "CathaCyc, a metabolic pathway database built from Catharanthus roseus RNA-Seq data", *Plant and Cell Physiology*, årg. 54, nr 5, s. 673–685, 2013. DOI: 10.1093/pcp/pct039.

[19] T. Sajed, A. Marcu, M. Ramirez, A. Pon, A. C. Guo m. fl., "ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli", *Nucleic Acids Research*, årg. 44, nr D1, s. D495–D501, 2015. DOI: 10.1093/nar/gkv1060.

[20] D. Wishart, Y. Djoumbou, A. Marcu, A. Guo, K. Liang m. fl., "HMDB 4.0: The human metabolome database for 2018", *Nucleic Acids Research*, årg. 46, 2017. DOI: 10.1093/nar/gkx1089.

[21] N. Sakurai, T. Ara, Y. Ogata, R. Sano, T. Ohno m. fl., "KaPPA-View4: a metabolic pathway database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data", *Nucleic acids research*, årg. 39, nr Database issue, s. D677–84, 2011. DOI: 10.1093/nar/gkq989.

[22] D. Cotter, E. Fahy och S. Subramaniam, "Tenth Annual LIPID Metabolites and Pathways Strategy (LIPID MAPS) Meeting", *Clinical Lipidology*, årg. 8, s. 403–405, 2013. DOI: 10.2217/clp.13.43.

[23] C. M. Andorf, E. K. Cannon, I. Portwood John L., J. M. Gardiner, L. C. Harper m. fl., "MaizeGDB update: new tools, data and interface for the maize model organism database", *Nucleic Acids Research*, årg. 44, nr D1, s. D1195–D1201, 2015. DOI: 10.1093/nar/gkv1007.

[24] H. Ginsburg och A. Mohamed, "Malaria Parasite Metabolic Pathways (MPMP) Upgraded with Targeted Chemical Compounds", *Trends in parasitology*, årg. 32, nr 1, 2015. DOI: 10.1016/j.pt.2015.10.003.

[25] M. Ganter, T. Bernard, S. Moretti, J. Stelling och M. Pagni, ”MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks”, *Bioinformatics*, årg. 29, nr 6, s. 815–816, 2013. DOI: 10.1093/bioinformatics/btt036.

[26] M. Sugimoto, S. Ikeda, K. Niigata, M. Tomita, H. Sato och T. Soga, ”MMMDB: Mouse Multiple Tissue Metabolome Database”, *Nucleic Acids Research*, årg. 40, nr D1, s. D809–D814, 2011. DOI: 10.1093/nar/gkr1170.

[27] ModelSEED, *ModelSEED*, https://modelseed.org/about/version, Online; 11 Maj 2021.

[28] A. V. Evsikov, M. E. Dolan, M. P. Genrich, E. Patek och C. J. Bult, ”MouseCyc: a curated biochemical pathways database for the laboratory mouse”, *Genome Biology*, årg. 10, nr R84, 2009. DOI: 10.1186/gb-2009-10-8-r84.

[29] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur m. fl., ”Pathway Commons, a web resource for biological pathway data”, *Nucleic acids research*, årg. 39, nr Database issue, s. D685–D690, 2011. DOI: 10.1093/nar/gkq1039.

[30] PMN, *Recent News*, https://plantcyc.org/about/news?page=4, Online; 11 Maj 2021.

[31] Reactome, *What is Reactome ?*, https://reactome.org/what-is-reactome, Online; 11 Maj 2021.

[32] U. Wittig, R. Kania, M. Golebiewski, M. Rey, L. Shi m. fl., ”SABIO-RK—database for biochemical reaction kinetics”, *Nucleic Acids Research*, årg. 40, nr D1, s. D790–D796, 2011. DOI: 10.1093/nar/gkr1046.

[33] H. Kim, J. Shin, E. Kim, H. Kim, S. Hwang m. fl., ”YeastNet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae”, *Nucleic Acids Research*, årg. 42, nr D1, s. D731–D736, 2013. DOI: 10.1093/nar/gkt981.

[34] M. Ramirez-Gaona, A. Marcu, A. Pon, A. Guo, T. Sajed m. fl., ”YMDB 2.0: a significantly expanded version of the yeast metabolome database”, *Nucleic acids research*, årg. 45, nr D1, s. D440–D445, 2017. DOI: 10.1093/nar/gkw1058.

[35] G. Bader, M. Cary och C. Sander, ”Pathguide: a pathway resource list.”, *Nucleic Acids Res.*, årg. 34, nr Database issue, s. D504–D506, 2006. DOI: 10.1093/nar/gkj126.

[36] D. Machado, S. Andrejev, M. Tramontano och K. R. Patil, ”Fast automated reconstruction of genome scale metabolic models for microbial species and communities”, *Nucleic Acids Res*, årg. 46, nr 15, s. 7542–7553, 2018. DOI: 10.1093/nar/gky537.

[37] F. Büchel, N. Rodriguez, N. Swainston, C. Wrzodek, T. Czauderna m. fl., ”Path2Models: large-scale generation of computational models from biochemical pathway maps”, *BMC Syst Biol*, årg. 7, nr 116, 2013. DOI: 10.1186/1752-0509-7-116.

[38] S. Magnúsdóttir, A. Heinken, L. Kutt, D. Ravcheev, E. Bauer m. fl., "Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota", *Nat Biotechnol*, årg. 35, nr 1, s. 81–89, 2017. DOI: 10.1038/nbt.3703.

[39] BioModels, https://www.ebi.ac.uk/biomodels/, Online; 6 Maj 2021.

[40] Systems Biology Research Group, *BiGG Models*, http://bigg.ucsd.edu, Online; 6 Maj 2021.

[41] KEGG, *KEGG: Kyoto Encyclopedia of Genes and Genomes*, https://www.genome.jp/kegg/, Online; 6 Maj 2021.

[42] M. Glont, T. Nguyen, M. Graesslin, R. Hälke, R. Ali m. fl., "BioModels: expanding horizons to include more modelling approaches and formats", *Nucleic acids research*, årg. 46, nr D1, s. D1248–D1253, 2018. DOI: 10.1093/nar/gkx1023.

[43] BioModels, *Path2Models project page*, https://www.ebi.ac.uk/biomodels/path2models, Online; 6 Maj 2021.

[44] R. van Heck, M. Ganter, V. Martins Dos Santos och J. Stelling, "Efficient Reconstruction of Predictive Consensus Metabolic Network Models", *PLoS Comput Biol.*, årg. 12, nr 8, e1005085, 2016. DOI: 10.1371/journal.pcbi.1005085.

[45] N. Pham, R. van Heck, J. van Dam, P. Schaap, E. Saccenti och M. Suarez-Diez, "Consistency, Inconsistency, and Ambiguity of Metabolite Names in Biochemical Databases Used for Genome-Scale Metabolic Modelling", *Metabolites*, årg. 9, nr 2, s. 28, 2019. DOI: 10.3390/metabo9020028.

[46] R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew och J. Nielsen, "The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum", *PLoS Comput. Biol.*, årg. 9, nr 3, e1002980, 2013. DOI: 10.1371/journal.pcbi.1002980.

[47] M. Kanehisa, "The KEGG database", i *'In Silico' Simulation of Biological Processes: Novartis Foundation Symposium 247*, G. Bock och J. A. Goode, utg., Hoboken, NJ, USA: Wiley Online Library, 2002, s. 91–103.

[48] A. Ravikrishnan och K. Raman, "Critical assessment of genome-scale metabolic networks: the need for a unified standard", *Brief Bioinform*, årg. 16, nr 6, s. 1057–68, 2015. DOI: 10.1093/bib/bbv003.

[49] M. A. Carey, A. Dräger, M. E. Beber, J. A. Papin och J. T. Yurkovich, "Community standards to facilitate development and address challenges in metabolic modeling", *Mol Syst Biol*, årg. 16, nr 8, e9235, 2020. DOI: 10.15252/msb.20199235.

[50] M. Hucka, A. Finney, B. Bornstein, S. Keating, B. Shapiro m. fl., "Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project", *Syst Biol (Stevenage)*, årg. 1, nr 1, s. 41–53, 2004. DOI: 10.1049/sb:20045008.

[51] C. Lieven, M. E. Beber, B. G. Olivier, F. T. Bergmann, P. Babaei m. fl., "Memote: A community driven effort towards a standardized genome-scale metabolic model test suite", *Nat Biotechnol*, årg. 38, s. 272–276, 2013. DOI: 10.1038/s41587-020-0446-y.

[52] I. Thiele och B. Ø. Palsson, "Reconstruction annotation jamborees: a community approach to systems biology", *Molecular Systems Biology*, årg. 6, nr 1, s. 361, 2010. DOI: https://doi.org/10.1038/msb.2010.15.

[53] Svensk Nationell Datatjänst, *Vad innebär FAIR Data?*, https://snd.gu.se/sv/beskriv-och-dela-data/vad-innebar-fair-data, Online; 12 Februari 2021.

[54] C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan m. fl., "NCBI Taxonomy: a comprehensive update on curation, resources and tools", *Database*, årg. 2020, 2020. DOI: 10.1093/database/baaa062.

[55] E. W. Sayers, M. Cavanaugh, K. Clark, J. Ostell, K. D. Pruitt och I. Karsch-Mizrachi, "GenBank", *Nucleic Acids Research*, årg. 47, nr D1, s. D94–D99, 2018. DOI: 10.1093/nar/gky989.

[56] R. S. Malik-Sheriff, M. Glont, T. V. N. Nguyen, K. Tiwari, M. G. Roberts m. fl., "BioModels — 15 years of sharing computational models in life science", *Nucleic Acids Research*, årg. 48, nr D1, s. D407–D415, 2020. DOI: 10.1093/nar/gkz1055.

[57] M. Glont, T. V. N. Nguyen, M. Graesslin, R. Hälke, R. Ali m. fl., "BioModels: expanding horizons to include more modelling approaches and formats", *Nucleic Acids Research*, årg. 46, nr D1, s. D1248–D1253, 2018. DOI: 10.1093/nar/gkx1023.

[58] NCBI National Center for Biotechnology Information, *NCBI ID Converter API*, https://www.ncbi.nlm.nih.gov/pmc/tools/id-converter-api/, Online; 9 Maj 2021.

[59] NCBI, *NCBI FTP Service*, https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/, Online; 9 Maj 2021.

[60] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang m. fl., "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks", *Genome Research*, årg. 13, nr 11, s. 2498–2504, 2003. DOI: 10.1101/gr.1239303.

[61] Lucid, *The intelligent diagramming application for every team*, https://lucid.co/product/lucidchart, Online; 9 Maj 2021.

[62] M. Denscombe, *The Good Research Guide*, 5. utg. Open University press, 2014, kap. 12.

[63] S. Brinkmann, *Qualitative Interviewing*. Oxford University Press, 2013, kap. 2.

[64] J. L. Reed, T. Vo, C. Schilling och P. BO, "An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)", *Genome Biology*, årg. 4, nr 9, 2003. DOI: 10.1186/gb-2003-4-9-r54.

A

Appendix 1

A.1 Intervjuer

A.1.1 Intervjufrågor

- What are you working on at the moment?
- What is the goal with your work?
- Why did you agree to this interview?
- When hearing what our project is about, what were your thoughts?
- Do you feel that there is a value to this research?
- What aspect of GEMs have you worked with before?
- Do you have any future plans of work or research within this field?
- As we introduced, the main focus and problem of our project is that there is no place or tool where for example model inheritance is gathered in an convenient way. Has this affected you in your work? If so, how?
- Regarding your research, do you recognize that our project would have helped you in your work? In case yes, in which way would it have helped you?
- If not model inheritance and reuse, what annotation parameter do you feel are of most importance?
- As mentioned, our goal is to enrich the existing annotation parameters and let this be a starting point of what we call, The Tree of GEMs. (Some kind of visualisation tool, which would be available at Metabolic Atlas.) Is there a specific question or problem you feel this would give answer to or solve?
- Are there any new questions this tool would evoke that has previously not been thought of?
- If this tool was available, would you use it for something specific? If so, what?
- Do you feel GEMs in general is part of a bigger picture?
- Would this project change you thoughts about this in any way?
- Is there any specific aspect that you would like us to focus on in our project and bachelors thesis?
- Is there any aspect you feel we have overlooked or missed to talk about? If so, feel free to raise any points you feel are important.

Avdelningen för Infrastrukturer inom Institutionen för Biologi och Bioteknik
CHALMERS TEKNISKA HÖGSKOLA
Göteborg, Sverige
www.chalmers.se

CHALMERS