

DATABASER
ADIT – TDDD74
Labbkompendium - Projekt

Institutionen för datavetenskap (IDA), Linköpings universitet

Innehåll

Översiktlig information	3
Projektlaboration: Bioinformatik	4
Delprojekt: Genomdatabas	7
Delprojekt: Proteindatabas	11
Delprojekt: Enzymsystem	16
Delprojekt: Biologiska reglersystem	20

Översiktlig information

För varje laboration skall det lämnas in en laborationsrapport.

Vänligen skriv ut din laborationsrapport på papper och kontrollera att den är lätt att läsa och förstå. Skicka inte in labbrapporter via mail om ni inte uttryckligen blir ombedda av er laborationshandledare att göra detta.

På kursens hemsida kan du hitta mer information och dokument gällande laborationerna.

Projektlaboration: Bioinformatik

Bakgrund

Informatik – konsten att lagra, förmedla, och presentera information – blir mer och mer viktigt inom biovetenskaperna, allteftersom forskarna gör nya rön. Datamängderna ökar mycket snabbt, och det ställs stora krav på de databassystem som ska underlätta forskningen i framtiden. En del av de system som finns idag har du säkert redan kommit i kontakt med.

Många av de stora biodatabaserna använder andra typer av databassystem än de du har kommit i kontakt med i kursen (relationsdatabaser). Ett skäl är bland annat att relationer inte är ett bra sätt att lagra all information som förekommer här (sekvensdata är ett typiskt fall – att lagra går bra, men det finns inte mycket support för sökning) – men däremot kan vissa bitar mycket väl hanteras av relationsdatabaser. Ofta får man kombinera relationer med andra tekniker, såsom olika typer av objektorienterade system och diverse specialprogramvara. I denna projektlaboration ska vi dock titta på de bitar som går bra att lagra i enbart relationsdatabaser.

Översikt över projektet

Under projektlaborationen ska du (din grupp) lösa ett litet bioinformatikproblem genom att skapa en relationsdatabas och fylla den med data. Andra grupper kommer att lösa andra, besläktade problem, och mot slutet av projektet kommer ni att kunna koppla ihop systemen till ett, och då kunna lösa mer komplexa uppgifter. Tanken är att ni på detta sätt ska kunna åstadkomma ett lite mer ”verkligt” system, som kan lösa svårare frågor. Det finns fyra olika delprojektuppgifter (biologiska regelsystem, enzymsystem, genomdatabas och proteindatabas), som vardera är lämplig för 2 personer.

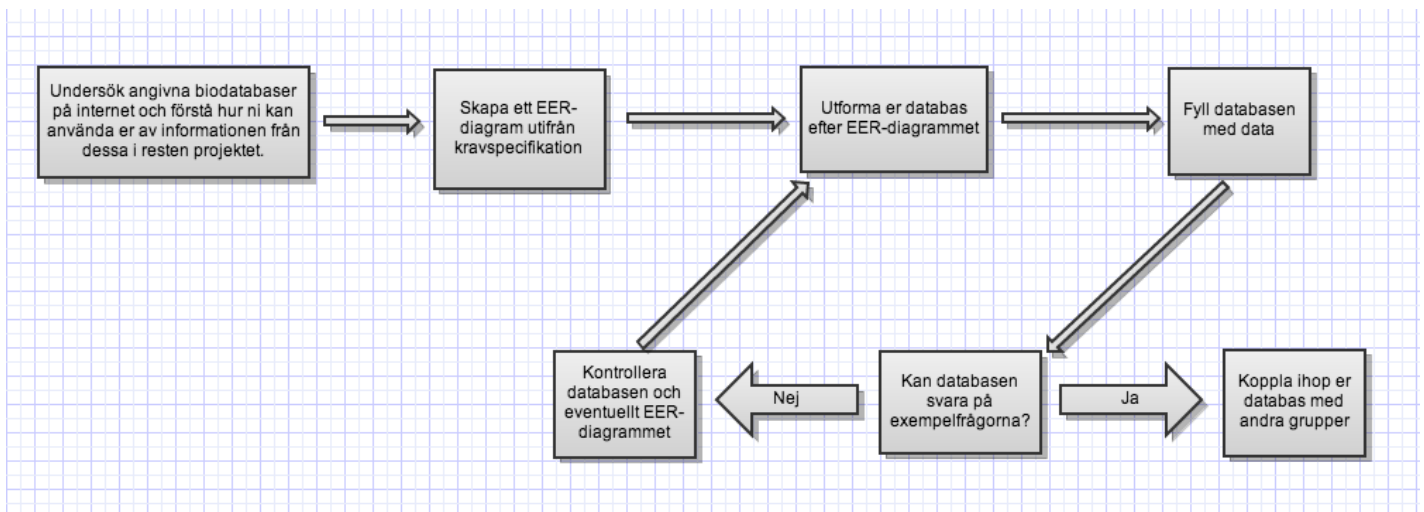
Varje grupp kommer att få en projekthandledning, utformat som en fallstudie. Ni kommer att få en viss bit biologisk data, och ska utforma er databas så att den klarar av att representera denna exempeldata på ett vettigt sätt. Ni kommer också att få några exempel på frågor om datamängden som databasen bör kunna svara på. Detta för att ni ska ha några riktlinjer i utformningen. Vi hoppas att du med denna projektlaboration ska få en lite bättre insikt i hur informationssystem som hanterar biologisk data fungerar, och en försmak av de problem man ställs inför när man jobbar med dem.

Mål

Genom projektet skall du uppnå följande mål:

- använda riktiga databaser och förstå hur situationen ser ut i verkligheten
- lära dig att söka i databaser som används i verkligheten
- lära dig att skapa EER-diagram
- lära dig översättning av EER till relationsschema
- lära dig att ställa frågor med hjälp av SQL
- lära dig vikten av terminologi och samarbete mellan olika databasdesigners

Översiktligt genomförande



Redovisning

Se kursens hemsida för deadlines.

1. Delprojektgrupp av 2 personer:

- lämna in och diskutera EER-diagrammet av er lösning i samband med en lektion samt vid projektredovisningen
- lämna in och diskutera databasschemat i samband med projektredovisningen
- lämna in SQL-frågor och resultat av de frågor som finns under rubriken 'Exempel på vanliga frågor' i samband med projektredovisningen

2. Projektgrupp bestående av 4 delprojektgrupper

- koppla ihop delprojektdatabaserna och lämna in SQL-frågor och resultat av de frågor som finns under rubriken 'Exempel på vanliga frågor' i samband med projektredovisningen

Kommentar

Som ni säkert har märkt så presenteras den information man får fram ur biodatabaserna oftast som textstycken, ibland väldigt långa sådana med en massa information i mer eller mindre begripligt format. Skälet är att biodatabaserna oftast lagrar sin data som textfiler, med en viss syntax som gör det möjligt att tolka informationen automatiskt i någon mån.

Ni kommer att använda en relationsdatabas, och därmed lagra informationen annorlunda, vilket i vissa fall är bättre och i vissa fall sämre. I detta fall kommer ni att märka att er konstruktion blir mindre flexibel, detta eftersom ni kommer att klassificera all data (som olika entitetstyper i EER-modellen). Textformatet är mer flexibelt i detta avseende, men är därför också svårare att arbeta med då det gäller automatisk tolkning och datakonvertering.

Det underliggande problemet är förstås att biologisk data ofta är svår att klassificera. Vi tenderar att göra för grova indelningar av biologin, vilket leder till att vi snabbt hittar undantag som inte passar i vår modell. Detta är ett svårt, övergripande problem i bioinformatiken. Grafiska presentationer som signalvägar, kromosomkartor, genssekvenser, 3D-strukturer och andra diagram är önskvärda, men kräver förstås en hel del merarbete. Ni

skulle i princip kunna skriva lika snygga grafiska applikationer för att presentera data ur er databas, men det är förstås för mycket arbete för denna lilla projektkurs.

Det är lite tråkigt: ni kommer att märka att det är svårt att få "snygga" svar ur er databas med en enkel SQL-tolk. Delvis beror detta på att ni i modelleringen och implementationen har delat upp och strukturerat om informationen. Ni är, till exempel, vana att se information om en signalväg som ett antal "påverkanspilar" mellan olika ämnen, kanske med siffror som anger förstärkningar och dylikt. I er databas kommer samma information att hamna i en massa tabeller. Men – och det är viktigt – så länge ni har lyckats få med all information, utan att förenkla bort en del aspekter av den, så har ni ett fullt fungerande informationssystem. Ni har då lyckats med huvuduppgiften, att lagra informationen - det är bara lite kosmetika som saknas.

Tack till ...

Denna projektuppgift har utvecklats av Roland Nilsson i samarbete med Maria Kolak, Lena Eklund, Almut Herzog och Patrick Lambrix och sedan reviderats av Elin Larsson och He Tan, Martin Gollvik och Rebecka Melin Rydfalk samt Patrick Lambrix.

Delprojekt: Genomdatabas

Syfte

Denna projektlaboration går ut på att modellera data om gener och nukleinsyror och kringliggande information. Målet är att skapa en egen genomdatabas i en relationsdatabas. En ganska informell kravspecifikation följer längre ner.

Bakgrund

Existerande databaser

Innan ni börjar med er egen databas ska ni börja med att utforska de befintliga databaser och informationssystem som finns på nätet. Detta för att ni få perspektiv på ert projekt och förstå hur informationen ni ska lagra ser ut samt lära er att hitta denna. Gendatabaser är nog det vanligaste på nätet, och det finns numera en uppsjö olika: NCBI/Genbank, DDBJ, EMBL-EBI, m.fl. Det vanligaste är att man lagrar sekvens, locusinformation, sekvenseringsinformation samt länkar till litteratur och proteiner. Länkar till databaserna finns på kursens hemsida.

- NCBI, National Center for Biotechnology Information, är USAs största bioinformatikresurs, och har en uppsjö olika databastjänster. Entrez är ett integrerat söksystem på NCBI som används flitigt. Det har en bra korslänkad struktur som omfattar Genbank, populationsgenetiska data, sekvens och 3D-struktur hos proteiner, och litteratordata.
- DDBJ (DNA Databank of Japan) och EMBL (European Molecular Biology Laboratory) har motsvarande databaser över forskningsresultat från Japan resp. Europa. NCBI, DDBJ och EMBL tillsammans utgör det som brukar kallas för International Nucleotide Sequence Database Collaboration, och utbyter ständigt information med varandra.

Sökuppgift

Bekanta dig med söksystemet Entrez på NCBI, på <http://www.ncbi.nlm.nih.gov/Entrez/>. Prova att leta fram genen för enzymet aconitase, som ni sedan kan välja att modellera i ert projekt. Börja med att välja "Nucleotide" som databas, och sök på nyckeln "NM_002197". Det här är en human gen för aconitase. Titta noga på de datablad ni får fram ur Entrez – jämför med den information som finns i kravspecifikationen och fundera på hur ni skulle modellera denna data. Hur hänger informationen ihop? Detta kan ge er tips om hur ni kan lagra informationen i er databas. Prova att göra fritextsökningar på exempelvis "aconitase". Du märker att det finns väldigt mycket material från många olika organismer och sammanhang – det är konst att söka tillräckligt precist. Sök därefter data från de andra exemplen nedan.

Titta sedan även på DDBJ, <http://www.ddbj.nig.ac.jp>. Klicka t.ex. på "Search/Analysis", sedan på "getentry" och sök på accession number DQ496106. I stället för "getentry" kan man också använda "ARSA" och söka på aconitase. Vilket accession number en gen har i olika databaser kan man hitta i GeneCards. Om ni söker på aconitase i GeneCards och tittar på ACO1 så får ni id nummer i andra databaser. Bekanta dig med söksystemet på EMBL, på <http://www.ebi.ac.uk/services>.

Denna uppgift behöver ej redovisas.

Kravspecifikation

Vi ska utgå från några gener som ni förhoppningsvis redan har stött på. Dessa exempel får utgöra kravspecifikation för ert system: er databas ska designas så att den klarar av att modellera den data som ingår i exemplen. Ni behöver inte lagra data om alla exemplen i er databas men databasen ska klara av att modellera den. I er databas ska ni lagra tillräckligt mycket data så att det går att se att databasen fungerar då ni ställer frågor till den.

Nedan följer en lista med information som ni ska klara av att lagra i er databas samt förslag på ytterligare information som ni kan integrera i er databas (valfritt). Fundera även kring om det finns annan information som ni vill kunna lagra och lägg gärna till detta i databasen om ni har möjlighet.

Information som databasen ska kunna hantera:

- NCBI GI-nummer. GI-numret hittar du strax under menyraden i sökningar i NCBI's Entrez. Det finns andra kandidatnycklar i NCBI, men GI har den trevliga egenskapen att det råkar vara ett heltal.
- Namn. Inte helt trivialt då det gäller biologiska strukturer, det finns alltid synonymer. Konstruera därför databasen så att en gen med ett visst GI-nummer kan ha flera olika namn.
- Versionshantering – kan vara ett knepigt problem ibland. Tänk på att de data kring gener som ni kommer att lagra är forskningsresultat, med en viss osäkerhet inbyggd – inte universella sanningar. Det är alltså troligt att nya, förhoppningsvis mer exakta data dyker upp. Ni ska kunna lagra olika versioner parallellt, för man vill troligen också ha kvar (och kunna hitta) äldre data. Ibland kan ju också informationen representera olika samtida åsikter om något.
- Sekvensinformation: datum för sekvensering, upphovsman, sekvensens längd. Själva bassekvensen bör ni nog strunta i. Den är besvärlig att lagra (bör hanteras som en s.k. BLOB i relationsdatabaser), och ni kommer inte att kunna göra några vettiga sökningar på den i alla fall.
- Olika klassificeringar: DNA eller RNA? Kromosomalt eller mitokondriellt? Utkast till sekvens, eller en reviderad upplaga?
- Heterobindningar, alltså andra ämnen som kan koppla till olika "sites" på syran på ett eller annat sätt, t.ex. metyleringar, bindningar med reglerande enzym, etc. Vanligen har ju såna modifieringar en reglerande funktion. Gruppen som gör enzymsystemdatabasen modellerar också mindre molekyler, länka gärna till dessa.
- Länka till relaterade artiklar om genen och/eller nukleinsyran. Lagra författare, rubrik och MedLine-nummer eller motsvarande.
- Länka till de protein/proteiner som genen uttrycker, lämpligen till en entitet hos gruppen som gör proteindatabasen.

Förslag på ytterligare information att lagra:

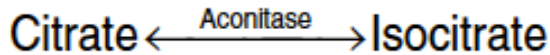
- Polymorfismer. Precis samma gen kan ju ha olika möjliga sekvenser.
- Gener förekommer ju ibland i kluster, operoner, och andra grupperingar. Den informationen är viktig bl.a. i exemplet om Lac-operonet (prokaryot) nedan.
- Övriga fysikaliska/kemiska egenskaper kring nukleinsyorna. Det är dock ganska sparsamt med sådan information på nukleindatabaserna.

Exempel att lagra

Minst ett av dessa exempel skall lagras i er databas. Prata igenom med de andra grupperna så att det går att göra några sökningar i den sammansatta databasen när allt är lagrat och klart.

Aconitase

Aconitase är ett av enzymen i TCA-cykeln. Det utför en isomerisering av citrat enligt nedan:



Aconitase är ett s.k. isoenzym, och har två kodande gener, Aco1 (GI:8659554) och Aco2 (GI:46411160). Aco1 kodar för ett cytosoliskt protein medan Aco2 kodar ett mitokondriellt, men båda generna är kromosomala.

Aco1 (och många besläktade gener) råkar också vara en s.k. IRE, "Iron-responsive element", vilket betyder att genen är inblandad i en regleringsmekanism på translationell nivå för ferritin och transferrinreceptorn. Därför får ni många träffar kring IRE och IREBP (IRE-binding protein, regulatoriskt protein) om ni söker på aconitase.

β-adrenalinreceptorn

En adrenoreceptor, ansvarig för detektion av adrenalin (epineprin). Två varianter, β-1 (GI:110349783) och β-2 (GI:6636495), har liknande struktur men något olika fysiologiska roller. Vill ni ha mer info om dessa proteiner i sitt biologiska sammanhang, så rekommenderas Zubay: Biochemistry s.328 resp. s.473.

Sök gärna vidare i databaserna efter besläktade gener, och fundera på hur ni skulle lagra och länka till dessa i er databas.

Insulin

Insulin uttrycks i människan i bukspottkörteln, i de s.k. "islets of Langerhans", en viss typ av cellkluster i detta organ. Signalmolekylen börjar sitt liv som preproinsulin (en förstadie till proinsulin).

För data om genen för humant preproinsulin, se t.ex. GenBank GI: 109148525.

MAPK

MAPK-familjen, "mitogen-activated protein kinases" (också känd som ERK-familjen, "extracellular-signal-regulated kinases") är en grupp enzymer med en nyckelroll i det regleringsystem som kallas MAPK-kaskaden.

Ett exempel på ett MAPK är humant MAPK3 (eller p44MAPK), som bland annat medierar insulinsignalen till genreglering. Locus för MAPK3 är 16p11.2. Dessa gener har också visat sig vara potentiella "targets" för cancerogena ämnen.

Se vidare under GI: 31220. 9

Lac-operonet: prokaryot genreglering

Ett exempel på genreglering finns i Lac-operonet i E.Coli. Operonet uttrycker enzymet β-galaktosidas, som spjälkar laktos till glukos, samt ett par relaterade enzymer.

Lac-i kodar för en repressor som i frånvaro av laktos binder till operatorlocus Lac-o och därmed hindrar uttryck av strukturproteinerna Lac-z, Lac-y, och Lac-a. Det finns också en

glukosberoende reglering genom repressorproteinet CAP, som binder till operatorlocus vid höga glukosnivåer. Operatorn har en karaktäristisk palindromsekvens (dyadsymmetri), som ger nukleinsyrans 3D-struktur hög affinitet mot det symmetriska CAP-proteinet.

Data om Lac-operonet finns bl.a. i GenBank GI: 146575, en ganska stor klump information.

Notera operatorregionen vid 1246...1266 samt CAP-bindande site vid 1162...1199. Se också Zubay, s.887-890.

Exempel på vanliga frågor

Som en ytterligare guide, så kan det vara bra att anpassa ert system så att det klarar följande vanliga frågor:

Givet en gen X:

- Vilken organism kommer gen X ifrån? Vilken vävnad?
- Var sitter gen X (locus-information) ?
- Vilka reglermekanismer finns? Är andra gener inblandade i regleringen?

Då ni har kopplat ihop er databas med de övriga delprojekten bör det också gå att ställa mer övergripande frågor som exempelvis

- Vilka proteiner uttrycker genen X?
- Vilka proteiner m.m. interagerar med gen X i regleringen? Vilka gener kodar för dessa proteiner?
- Vilka enzysystem deltar uttryckt protein i, alt. var i cellen finns det uttryckta proteinet?

Referenser

Zubay, Geoffrey L: "Biochemistry", 1998. The McGraw-Hill Companies, Inc. ISBN: 0697219003

Hancock, John T: "Cell Signaling", 1997. Addison Wesley Longman Ltd. ISBN 0582312671

Delprojekt: Proteindatabas

Syfte

Denna projektlaboration går ut på att modellera proteiner, deras egenskaper och annan relaterad information. Målet är att skapa en egen proteindatabas i en relationsdatabas. En ganska informell kravspecifikation följer längre ner.

Bakgrund

Existerande databaser

Innan ni börjar med er egen databas ska ni börja med att utforska de befintliga databaser och informationssystem som finns på nätet. Detta för att ni få perspektiv på ert projekt och förstå hur informationen ni ska lagra ser ut samt lära er att hitta denna. Det finns en hel del proteindatabaser: PDB, KEGG, NCBI/Entrez, Swiss-prot, m.fl. Länkar till databaserna finns på kursens hemsida.

- PDB, Protein Data Bank, är en databas med 3D-strukturdata om proteiner. Den data ni hittar här kommer vanligen från röntgenkristallografi eller NMR, men det kan också hända att strukturerna är teoretiskt framtagna med molekylfysiska simuleringar.
- NCBI, National Center for Biotechnology Information, är USAs största bioinformatikresurs, och har en uppsjö olika databastjänster. Entrez är ett integrerat söksystem på NCBI som används flitigt. Det har en bra korslänkad struktur som omfattar gensekvenser, genomkartor, populationsgenetiska data, sekvens och 3Dstruktur hos proteiner, och litteraturlänkdata.
- UniprotKB finns på servern <http://www.uniprot.org> tillsammans med många andra verktyg för proteinanalys. Databasen har många korsreferenser som kan hjälpa till att hitta bl.a. fysikaliska data om peptiderna.
- KEGG hanterar enzymsystem, och har ofta kopplingar till (och från) proteindatabaserna. Ni ska inte modellera enzymsystemen – det hanterar en annan projektgrupp – men det kan vara intressant att titta lite på enzymatiska proteiners biokemiska sammanhang.

Sökuppgift

Titta lite på NCBI/Entrez, PDB, UniProtKB och KEGG. Prova att leta fram enzymet aconitase i respektive databas. Databaserna finns på <http://www.ncbi.nlm.nih.gov/Entrez/>, <http://www.rcsb.org/pdb/>, <http://www.uniprot.org>, <http://www.genome.ad.jp/kegg/> respektive. Det är lättast att söka på GI-nummer i NCBI/Entrez (välj "Nucleotide" databas). (GI-numret för aconitase är 166158926.) Titta noga på de datablad ni får fram, och fundera på hur ni skulle modellera den data ni hittar. Om ni i PDB söker på EC-numret får ni ett flertal träffar där aconitase är i komplex med olika molekyler. Om ni klickar på varje namn så får ni mer information om de olika strukturerna. I UniprotKB kan ni t.ex. söka på aconitase human eller P21399. I KEGG kan ni skriva EC-numret (4.2.1.3) istället för aconitase. Prova att söka information om andra proteiner ur exemplen nedan. Jämför med den information som finns i kravspecifikationen och fundera på hur ni skulle rita EER-diagram.

Denna uppgift behöver ej redovisas.

Kravspecifikation

Vi ska utgå från några proteiner som ni förhoppningsvis redan har stött på. Dessa exempel får utgöra kravspecifikation för ert system: er databas ska designas så att den klarar av att modellera den data som ingår i exemplen. Ni behöver inte lagra data om alla exemplen i er databas men databasen ska klara av att modellera den. I er databas ska ni lagra tillräckligt mycket data så att det går att se att databasen fungerar då ni ställer frågor till den.

Nedan följer en lista med information som ni ska klara av att lagra i er databas samt förslag på ytterligare information som ni kan integrera i er databas (valfritt). Fundera även kring om det finns annan information som ni vill kunna lagra och lägg gärna till detta i databasen om ni har möjlighet.

Information som databasen ska kunna hantera:

- NCBI GI-nummer. GI-numret hittar du strax under menyraden i sökningar i NCBI's Entrez.
- Namn. Inte så lätt som man tror om man vill ta hänsyn till alla synonymer som finns; olika trivialnamn, olika (mer eller mindre) systematiska namn, etc. Konstruera därför databasen så att ett protein kan ha flera olika namn.
- Versionshantering – kan vara ett knepigt problem ibland. Tänk på att de data kring proteiner som ni kommer att lagra är forskningsresultat, med en viss osäkerhet inbyggd – inte universella sanningar. Det är alltså troligt att nya, förhoppningsvis mer exakta data dyker upp. Ni ska kunna lagra olika versioner parallellt, för man vill troligen också ha kvar äldre data. Ibland kan ju också informationen representera olika samtida åsikter om något.
- Klassificering. Enzym, strukturprotein, receptor? Kan vara en knepig distinktion men försök ändå att klassificera de proteiner som ni lagrar i er databas.
- Sekvensinformation: datum för sekvensering, upphovsman, sekvensens längd. Själva aminosyrasekvensen bör ni nog strunta i. Den är besvärlig att lagra (bör hanteras som en s.k. BLOB i relationsdatabaser), och ni kommer inte att kunna göra några vettiga sökningar på den i alla fall.
- Proteinkomplex (tertiärstruktur). I gängse terminologi kan ett protein ibland bestå av flera peptidkedjor, i komplex. Fundera över hur detta hanteras bäst.
- Heterobindningar, alltså andra ämnen som är kopplade till proteinet på ett eller annat sätt. Gruppen som gör enzysystem-databasen modellerar också mindre molekyler. Länka gärna till dessa då ni behöver beskriva dessa interaktioner.
- En del proteiner bildas ur andra ("precursors") genom diverse kovalenta modifieringar. Se till att ni kan hålla koll på denna "släktskap".
- Länka till relaterade artiklar om proteinet. Lagra författare, rubrik och MedLine-nummer eller motsvarande.

Förslag på ytterligare information att lagra:

- Sekundärstruktur. Lagra information om motivbildning (α -helix, β -flak, cysteinbryggor etc.). Lagra också aktiva/regulatoriska ytor och andra regioner med speciell betydelse (exempelvis transmembranhelixar i membranproteiner).
- Själva peptiderna deltar ju ibland som substrat/produkter i andra reaktioner. Det är kanske motiverat att låta en entitet "peptid" /motsv. vara en specialisering av en mer generell "substans"-entitet. Prata med någon ur enzysystem-gruppen, som jobbar med att modellera substanser i så fall.

- Peptidernas övriga fysikaliska/kemiska egenskaper: molekylvikt, isobestisk punkt, löslighet, etc.

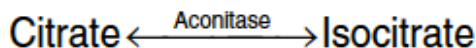
Mer info om dessa proteiner i sitt biologiska sammanhang hittar ni i Zubay: Biochemistry s.328 resp. s.473.

Exempel att lagra

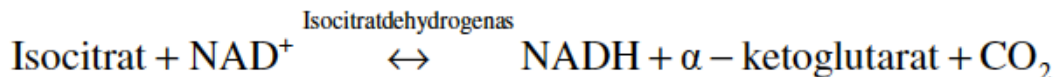
Minst ett av dessa exempel skall lagras i er databas. Prata igenom med de andra grupperna så att det går att göra några sökningar i den sammansatta databasen när allt är lagrat och klart.

TCA-cykeln: Aconitase och Isocitratdehydrogenas

Aconitase är ett enzym som utför en isomerisering av citrat enligt nedan:



Aconitase är ett s.k. iron-sulfur-protein: kofaktorer är Fe och S, som bildar ett 4Fe-4S kluster kopplat till peptiden. Aconitase tillhör gruppen hydro-lyaser. Detaljerad info finns i NCBI: sök i Entrez (välj "protein" eller "structure" som databas), ange GI-numret: 9802307. Enzymets EC-nummer är 4.2.1.3 om du vill titta på andra, liknande proteiner. Isocitratdehydrogenas katalyserar reaktionen. (Mer info hittar ni i Zubay, Geoffrey L.: Biochemistry s.331.)



Enzymet uppträder som homodimerer. EC-numret är 1.1.1.42.

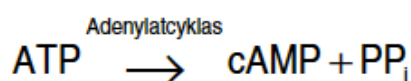
β-adrenalinreceptorn

En adrenoreceptor, ansvarig för detektion av adrenalin(ephineprin)signalen. Det finns inte (läs: jag har inte hittat) så bra strukturinformation kring denna receptor, även om den förekommer i litteraturen rätt ofta. Troligen är transmembranhelixarna belägna ungefär på intervallen 86-110, 120-140, 199-220, 243-264, 287-311, 323-348, samt 362-386. Aminoterminalen är extracellulärt belägen, regionen nära denna (400-418) utgör den regulatoriska ytan för inbindning av adrenalin. Mg²⁺ är en aktiverande kofaktor, men mekanismen är inte känd.

Två varianter, β-1 (GI: 110349783) och β-2 (GI:6636495), har liknande struktur men något olika fysiologiska roller. β-adrenalinreceptorn är normalt associerad med ett GS-protein vid den intracellulära Cterminalen. Vid inbindning till den regulatoriska ytan sker en konformationsförändring som får GS,α att lossna. Sök gärna efter mer information i proteindatabaserna.

Adenylatcyklas

Efter aktiveringen av β-adrenalinreceptorn kommer det frisatta GS,α att aktivera adenylatcyklas, som katalyserar reaktionen:



Adenylatcyklas ingår i många andra signalvägar, framförallt i kombination med olika Gproteiner. Humant adenylatcyklas från ett antal olika vävnader har sekvenserats och

strukturbestämts, och man har därmed hittat ett antal olika undertyper. Ex.vis. typ 9 (ADCY9) hittar ni på GI: 9802307 i Entrez/protein. Här hittar ni mycket information om olika domäner i proteinet, även om mycket är markerat "provisional" (inte validerat). Se också KEGG, EC 4.6.1.1.

Insulin

Insulin uttrycks i människan i bukspottkörteln, i de s.k. "islets of Langerhans", en viss typ av cellkluster i detta organ. Signalmolekylen börjar sitt liv som preproinsulin (en förstadie till proinsulin). De första omformningarna av peptiden kan sammanfattas som följer:

1. Preproinsulin uttrycks och transporteras till ER. En aa-sekvens vid N-terminalen i peptiden tjänar här som sorteringssignal.
2. En aa-sekvens klyvs i ER av från N-terminalen, och proinsulin bildas.
3. Proinsulin transporteras ut mot cellmembranet genom golgi och vesiceltrafiken. I processen klyvs ytterligare en aa-sekvens bort, nu från mitten av proteinet, och insulinhormonet har bildats. Detta peptidhormon har således två peptidkedjor, A och B, sammanhållna av disulfidbryggor.
4. Insulin går ut i blodkärlen och transporteras till målcellerna.

En sekvens för humant proinsulin med de olika regionerna utmärkta hittar ni på GI: 4557671.

MAPK

MAPK-familjen, "mitogen-activated protein kinases" (också känd som ERK-familjen, "extracellular-signal-regulated kinases") är en grupp enzymer med en nyckelroll i det reglersystem som kallas MAPK-kaskaden. Ett exempel på ett MAPK är humant MAPK3 (eller p44MAPK), som bland annat medierar insulinsignalen till genreglering. Detta enzym är en peptid-monomer på 379aa. Sök efter mer information på GI: 232066. 14 β -galaktosidas

Detta enzym förekommer i laktosmetabolismen i de flesta organismer. Vi ska titta på prokaryot β -galaktosidas från E.coli, som bör vara bekant från höstterminens mikrobiologikurs. Enzymet spjälkar disackarid-bindningen i laktos och producerar därmed två monosackarider, galaktos och glukos. Proteinets är relativt stort: en homotetramer med subenheter om 116 kD. Se GI: 114939. Där finner ni ett exempel på vikten av versionshantering: 1984 bestäms glu-461 vara aktiv site, 1992 argumenterar andra forskare för att detta är felaktigt och att glu-537 är den korrekta aminosyran.

Exempel på vanliga frågor

Som en ytterligare guide, så kan det vara bra att anpassa ert system så att det klarar följande vanliga frågor:

Givet ett protein X:

- Är proteinet X en monomer eller ett komplex? Om det är komplexbundet, vilka peptider är subenheter?
- Vilka andra ämnen interagerar med proteinet X? Hur (vilka "sites")?
- Vilka versioner av proteinet X finns (olika sekvenseringar, undertyper per vävnad, etc.)?
- Finns proteinet X i olika stadier, s.k. Precursors? Vilka är i så fall dessa?
- Finns det artiklar relaterade till proteinet X?
- Associerar proteinet X med andra proteiner till polymerer? (ex. kollagen)

- Innehåller dessa kanske även andra ämnen? (ex. peptidoglykan) ¹

Då ni har kopplat ihop er databas med de övriga delprojekten bör det också gå att ställa mer övergripande frågor som exempelvis.

- Vilka gener kodar för proteinerna som utgör ett enzym E?
- Vilka enzymssystem deltar proteinet X i?
- Vilka signalvägar förekommer proteinet X i?

Samt ta fram en massa information om dessa enzymer och gener, förstås. Men det sköter de andra grupperna.

Referenser

Zubay, Geoffrey L: "Biochemistry", 1998. The McGraw-Hill Companies, Inc. ISBN: 0697219003

Hancock, John T: "Cell Signaling", 1997. Addison Wesley Longman Ltd. ISBN 0582312671

¹ För strukturproteiner vore det ju användbart vore att kunna hitta associationer och bindningar i flera led. Detta kräver dock att man kan göra rekursiv sökning i databasen. Det går att göra i SQL3: se Elmasri & Navathe: Fundamentals of Database Systems, kap.22.1.3 (Ed5), om du är intresserad.

Delprojekt: Enzymsystem

Syfte

Denna projektlaboration går ut på att modellera enzymsystem, d.v.s. enzymer, enzymreaktioner, och alla ämnen som ingår i reaktionsmekanismerna på ett eller annat sätt. Målet är att skapa en databas som kan lagra information om dessa enzymsystem. En ganska informell kravspecifikation följer längre ner.

Bakgrund

Existerande databaser

Innan ni börjar med er egen databas ska ni börja med att utforska de befintliga databaser och informationssystem som finns på nätet. Detta för att ni få perspektiv på ert projekt och förstå hur informationen ni ska lagra ser ut samt lära er att hitta denna.

En användbar databas för er uppgift är KEGG (Kyoto Encyclopedia of Genes and Genomes). KEGG lagrar för närvarande mest data om metaboliska enzymer, men man har ambitionen att även lagra regulatoriska pathways (olika signalsystem alltså). I allmänhet så är informationssystemen på webben ganska dåliga på biokemiska signalvägar, men de blir bättre undan för undan. KEGG har också ett bra grafiskt gränssnitt (representation, tolkning av informationen) till databasen i form av klickbara metabolismkartor. I dessa kan man lätt – om man har lite biokemisk allmänbildning, alltså – leta reda på de enzymer som är inblandade i just de reaktioner man är intresserad av.

Även Swiss Institute of Bioinformatics (SIB) har enzymdatabaser, tillsammans med massor av verktyg för ”proteomics”, olika former av protein-analys. Slutligen är BRENDA en bra källa för bl.a. kinetikdata och annan kemisk information. Länkar till databaserna finns på kursens hemsida.

Uppgift

Öppna KEGG i en webbläsare. Adressen är <http://www.genome.ad.jp/kegg/>. Sök på ”aconitase”. Här får ni flera träffar. Längst ner på sidan hittar ni KEGG ENZYME. Klicka på EC-numret (4.2.1.3) och titta på hur informationen är lagrad. Jämför med informationen i kravspecifikationen. Fundera på hur ni skulle rita EER-diagram. För att se kartor över de metaboliska vägar aconitase deltar, använd KEGG PATHWAY (på första sidan) och sök på EC-numret. Sök även information om de andra exemplen nedan.

Titta också på enzymdatabaserna i Expasy (<http://www.expasy.ch/enzyme>) (enzymdata). Dessa databasverktyg är sammanlänkade och använder samma EC-nummer som KEGG för sina primärnycklar. Titta slutligen i BRENDA på <http://www.brenda-enzymes.info/>. Prova att söka i databasen genom att söka på enzymnamn eller samma EC-nummer som i KEGG.

Denna uppgift behöver ej redovisas.

Kravspecifikation

Vi ska utgå från några enzymer som ni förhoppningsvis redan har stött på. Dessa exempel får utgöra kravspecifikation för ert system: er databas ska designas så att den klarar av att modellera den data som ingår i exemplen. Ni behöver inte lagra data om alla exemplen i er

databas men databasen ska klara av att modellera den. I er databas ska ni lagra tillräckligt mycket data så att det går att se att databasen fungerar då ni ställer frågor till den.

Nedan följer en lista med information som ni ska klara av att lagra i er databas samt förslag på ytterligare information som ni kan integrera i er databas (valfritt). Fundera även kring om det finns annan information som ni vill kunna lagra och lägg gärna till detta i databasen om ni har möjlighet.

Information som databasen ska kunna hantera:

- EC-numren ur KEGG, de är en IUPAC-standard och används lite överallt. De är inte enkla heltal, så använd strängar.
- Namn. Inte så lätt som man tror om man vill ta hänsyn till alla synonymer som finns; olika trivialnamn, olika (mer eller mindre) systematiska namn, etc. Tyvärr har ju inte standardiseringen kommit så värst långt i detta område.
- Klassificering (tillhörighet i enzymfamilj), exempelvis lyaser, proteaser, etc. Familjer har ofta undergrupper, ibland i flera nivåer (som trädstrukturer).
- Det protein som utgör enzymet. Här bör ni helt enkelt länka till den grupp som gör proteindatabasen.
- Vilka substrat och produkter enzymet har (olika kemiska ämnen), för att beskriva en enzymatisk reaktion. Lagra även stökiometrin. Andra trevliga saker är affinitet för substrat, diverse kinetisk information, huruvida reaktionen är reversibel, o. dyl. Tänk på att ett enzym kan katalysera flera reaktioner ibland! (eller ska man kanske betrakta det som att ett protein då utgör flera enzym – har flera roller, så att säga...? Fundera.)
- Kofaktorer eller koenzym som enzymet interagerar med. Dessa kan förstås vara inhiberande såväl som aktiverade.
- Namn (kemiskt och/eller trivial) för mindre molekyler som ingår i reaktionerna. Samma problem som ovan. Summaformel, molekylvikt, löslighet, etc. är förstås andra attribut man kan spä på med om man vill.

Förslag på ytterligare information att lagra:

Reaktioner med makromolekyler (proteiner, nukleinsyror, polymerer) som substrat och produkt är lite knepiga. Ofta är det ju bara en liten del av molekylen som reagerar (exempelvis då ett restriktionsenzym klipper specifikt bakom en viss aminosyra), och det är i dessa fall relevant att veta vilket "site" på molekylen som deltar i reaktionen. Den informationen är bra att kunna lagra. Prata med grupperna som gör proteindatabasen eller gendatabasen och se om ni kan länka till deras "sites".

Allmänt

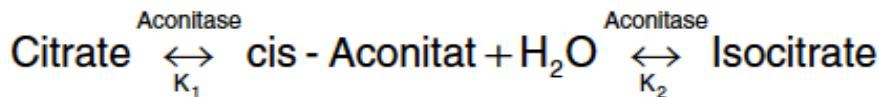
Lagra inte en massa data som har att göra med enzymernas proteinstruktur. Den grupp som arbetar med proteindatabasen kommer att göra detta ganska grundligt. Fokusera på reaktionerna. Några bekanta system som får utgöra exempeldata beskrivs nedan. Mer info hittar ni i Zubay, Geoffrey L.: Biochemistry s.328 resp. s.473.

Exempel att lagra

Minst ett av dessa exempel skall lagras i er databas. Prata igenom med de andra grupperna så att det går att göra några sökningar i den sammansatta databasen när allt är lagrat och klart.

TCA-Cykeln: Aconitase & Isocitratdehydrogenas

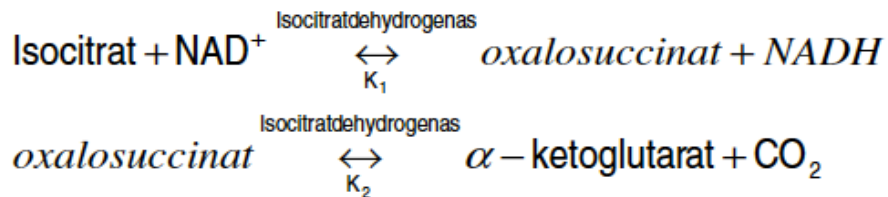
Aconitase är ett av enzymen i TCA-cykeln. Det utför en isomerisering av citrat enligt nedan:



Enzymet katalyserar alltså två reaktioner, båda reversibla. $K_1=0,07$ och $K_2=0,7$.

Aconitase är ett s.k. iron-sulfur-protein. Fe^{2+} är aktivator och hjälper till att bilda ett 4Fe-4S-kluster kopplat till enzymet. Mer information finns i KEGG, EC är 4.2.1.3.

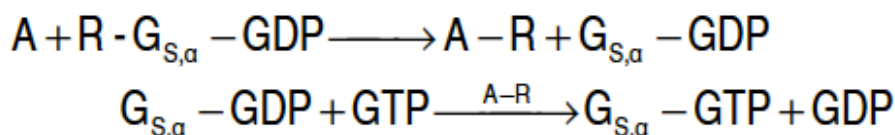
Nästa steg i TCA-cykeln, bildandet av α -ketoglutarat, är också ett viktigt reglerande steg. Enzymet isocitratdehydrogenas inhiberas av ATP och NADH, och aktiveras av NAD^+ och ATP (således stängs det av vid hög "energinivå" i cellen). Reaktionen är:



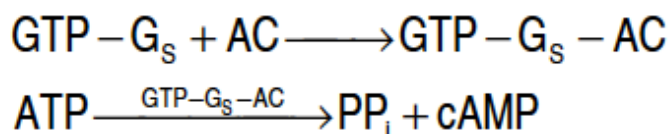
Där intermediären oxalosuccinat sitter bunden till enzymet. Sök på EC 1.1.1.42. Se också Zubay s.257.

β -adrenalinreceptorn och adenylatcyclas

Denna receptor är ansvarig för medieringen av adrenalinsignalen, som orsakar stora omställningar i cellen. Det första steget i denna cellaktivering är förstås själva inbindningen av adrenalin (A) till receptorn (R). Inbindningen inducerar en konformationsförändring som får ett G-protein att byta ut ett bundet GDP mot GTP, lossna från receptorkomplexet och diffundera mot ett närliggande adenylatcyclas:



Därefter binder G-proteinet till, och aktiverar därmed Adenylatcyclas, som katalyserar syntesen av cAMP:

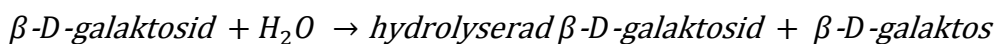


Vidare kommer cAMP att aktivera cytosoliska cAMP-beroende proteinkinaset, som senare fosforylerar en herrans massa proteiner för att ställa om cellen till "stressat tillstånd". Vi stannar dock vid dessa reaktioner.

β -galaktosidas

Detta enzym förekommer i laktosmetabolismen i de flesta organismer. Vi ska titta på prokaryot β -galaktosidas från E.coli, som bör vara bekant från höstterminens mikrobiologikurs. Enzymet spjälkar disackarid-bindningen i laktos och producerar därmed två monosackarider, galaktos och glukos. Enzymet är inte så känsligt för strukturvariationer i sitt substrat, och kan arbeta med en massa olika substratanaloger.

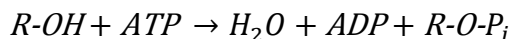
Den allmänna reaktionen är:



Allolaktos uppträder i reaktionen som en intermediär. EC är 3.2.1.23. Titta i BRENDA efter kinetiska parametrar.

MAPK

MAPK-familjen, "mitogen-activated protein kinases" (också känd som ERK-familjen, "extracellular-signal-regulated kinases") är en grupp enzymer med en nyckelroll i det reglersystem som kallas MAPK-kaskaden. MAPK-enzymerna är serin/threonin-kinaser, och tillhör EC-klassen 2.7.1: fosfotransferaser med OH som acceptor. De aktiveras själva av andra kinaser, och katalyserar då reaktionen



Där R är serin- eller threoninresidualer på målproteinerna. Vilka proteiner som fosforyleras beror förstås på vilket MAPK vi pratar om. Exempelvis fosforylerar humant MAPK3 (aka. ERK1 eller p44ERK1) proteinerna MAP2, MBP, och ELK1, och påverkar troligen cellcykeln. MAPK3 aktiveras genom insulinsignalen. Se också Zubay, s.216ff om fosforylering i allmänhet.

Exempel på vanliga frågor

Som en ytterligare guide, så kan det vara bra att anpassa ert system så att det klarar följande vanliga frågor:

- Givet ett ämne X:
 - Vilka (enzym)reaktioner producerar ämnet/förbrukar ämnet?
 - Vilka enzym inhiberas/aktiveras av detta ämne?
- Givet två ämnen, X och Y:
 - Vilka enzymer kan konvertera X till Y?²
- Givet ett enzym E:
 - Vilka reaktioner deltar E?

Då ni har kopplat ihop er databas med de övriga delprojekten bör det också gå att ställa mer övergripande frågor som exempelvis

- Vilket/vilka proteiner/proteinkomplex utgör E?
- Vilka gener kodar för proteinerna som utgör E?
- I vilka signalsystem ingår E?

Samt ta fram en massa information om dessa proteiner och gener, förstås. Men det sköter de andra grupperna.

Referenser

Zubay, Geoffrey L: "Biochemistry", 1998. The McGraw-Hill Companies, Inc. ISBN: 0697219003

Hancock, John T: "Cell Signaling", 1997. Addison Wesley Longman Ltd. ISBN 0582312671

² Ytterst användbart vore att kunna hitta hela pathways mellan två olika ämnen. Detta kräver dock att man kan göra rekursiv sökning i databasen. Det går att göra i SQL3: se Elmasri & Navathe: Fundamentals of Database Systems, kap.22.1.3 (Ed5), om du är intresserad. För den här uppgiften kan du i stället anta att pathways är 1, 2 eller 3 steg långa.

Delprojekt: Biologiska reglersystem

Syfte

Denna projektlaboration går ut på att modellera reglersystem och signalvägar i celler och organismer. Målet är att skapa en databas i en relationsdatabas som sedan kan kopplas till de genom-, protein- och enzymssystemdatabaser som de andra projektgrupperna gör. En ganska informell kravspecifikation för ert projekt följer längre ner. Observera att er uppgift endast är att modellera signalvägarna - detaljerna kring hur reaktionerna går till (stökiometri, kofaktorer, etc.) görs av gruppen som modellerar enzymssystem.

Bakgrund

Existerande databaser

Innan ni börjar med er egen databas ska ni börja med att utforska de befintliga databaser och informationssystem som finns på nätet. Detta för att ni få perspektiv på ert projekt och förstå hur informationen ni ska lagra ser ut samt lära er att hitta denna.

Det finns väldigt mycket biologisk data i olika informationssystem på www. I dagsläget är den största delen av denna information gen- och proteindata, men det finns varianter som uppmärksammar reglersystem. MetaCyc (BioCyc) är ett exempel. Den representerar signalvägar som kedjor av relationer, och innehåller information om de kemiska substanser som ingår och länkar till litteratur. Den presenterar till stor del data grafiskt, som diagram över signalvägar.

Också databasen KEGG (Kyoto Encyclopedia of Genes and Genomes) lagrar data om regulatoriska pathways (olika signalsystem alltså). Länkar till databaserna finns på kursens hemsida.

Uppgift

Öppna Metacyc på <http://www.metacyc.org/>. Sök på insulin. Då får ni upp länkar till pathways. Klicka på "Pathways" under "Search" (på första sidan) och sök på TCA cycle. Klicka på en instans för att få upp en karta över en signalväg. Man kan få mera eller mindre information genom att klicka på "More detail" och "Less detail".

En annan någorlunda relevant databas är KEGG, Kyoto Encyclopedia of Genes and Genomes, på <http://www.genome.jp/kegg/>. Här finns en massa information om "pathways", reaktionsvägar, men inte så mycket specifikt om reglermekanismer. Använd KEGG PATHWAY för att söka information.

Denna uppgift behöver ej redovisas.

Kravspecifikation

Vi ska utgå från några gener som ni förhoppningsvis redan har stött på. Dessa exempel får utgöra kravspecifikation för ert system: er databas ska designas så att den klarar av att modellera den data som ingår i exemplen. Ni behöver inte lagra data om alla exemplen i er databas men databasen ska klara av att modellera den. I er databas ska ni lagra tillräckligt mycket data så att det går att se att databasen fungerar då ni ställer frågor till den.

Nedan följer en lista med information som ni ska klara av att lagra i er databas samt förslag på ytterligare information som ni kan integrera i er databas (valfritt). Fundera även kring om det

finns annan information som ni vill kunna lagra och lägg gärna till detta i databasen om ni har möjlighet.

Information som databasen ska kunna hantera:

- Signaler: sändare, mottagare (olika substanser), information om signalen förstärks eller försvagas av ett ämne i överföringen, information om överföringsmekanismen.
- Länka till relaterade artiklar om signalmekanismen. Lagra författare, rubrik och MedLine-nummer eller motsvarande.
- Länka till motsvarande protein/proteiner, lämpligen till en entitet hos gruppen som gör proteindatabasen.

Förslag på ytterligare information att lagra

- Fundera på skillnaden mellan en signal och den underliggande kemiska reaktionen – är det någon skillnad? Gruppen som gör enzymdatabasen gör också en modellering över kemiska reaktioner. Det kan kanske vara vettigt att låta en signalmekanism vara en specialisering av en sådan kemisk reaktion?

Exempel att lagra

Minst ett av dessa exempel skall lagras i er databas. Prata igenom med de andra grupperna så att det går att göra några sökningar i den sammansatta databasen när allt är lagrat och klart.

TCA-cykeln

TCA-cykeln regleras huvudsakligen vid nyckelenzymerna Pyruvatdehydrogenas och Isocitratdehydrogenas. En bra översikt över regleringen hittar du i Zubay, G.L.: Biochemistry, s.341.

β -adrenalinreceptorn

En adrenoreceptor, ansvarig för detektion av adrenalin (epineprin).

Signalvägen är, i korthet:

1. Adrenalin binder in till β -adrenalinreceptorn och inducerar därmed en konformationsförändring som får ett till receptorn associerat GS-protein att släppa sin α -subenhet.
2. GS, α byter ett associerat GDP mot GTP och diffunderar längs membranet (2-dimensionell diffusion?) till ett närliggande adenylatcyclas som aktiveras.
3. Adenylatcyclas syntetiserar en mängd cAMP under sin aktiva tid, och stängs sedan av.
4. Den höjda cAMP-nivån i cellen aktiverar ett cAMP-beroende proteinkinase (cAPK), som medierar signalen vidare till cellrespons genom ett otal proteinfosoryleringar. Fosfodiesteras stänger så småningom av signalen genom att bryta ner cAMP och därmed stoppa cAPK.

Signalvägen beskrivs i Zubay s.473, och Hancock, "Cell Signaling", kap. 3 och 5. Läs noga och begrunda vilka moment i transduktionen som är signaler!

Insulin

Ett tredje exempel är insulinsignalen, som sammanfattas nedan. Insulin uttrycks i människan i bukspottkörteln, i de s.k. "islets of Langerhans", en viss typ av cellkluster i detta organ. Signalmolekylen börjar sitt liv som preproinsulin (en förstadie till proinsulin). En ungefärlig

beskrivning av signalvägen följer. Observera att många steg är osäkra: du kan säkert hitta andra, delvis motstridiga uppgifter.

1. Preproinsulin uttrycks och transporteras till ER. En aa-sekvens vid N-terminalen i peptiden tjänar här som sorteringssignal.
2. En aa-sekvens klyvs i ER av från N-terminalen, och proinsulin bildas därmed.
3. Proinsulin transporteras ut mot cellmembranet genom golgi och vesiceltrafiken. I processen klyvs ytterligare en aa-sekvens bort, nu från mitten av proteinet, och insulinhormonet har bildats. Detta peptidhormon har således två peptidkedjor, A och B, sammanhållna av disulfidbryggor.
4. Insulin går ut i blodkärlen och transporteras till målcellerna.
5. Insulin binder till insulinreceptorn vid dess α -subenheter, och inducerar en autofosforylering i receptorns transmembrana β -subenheter. Detta håller receptorn aktiv även om insulinmolekylen skulle lossna.
6. Den nu aktiva receptorn har tyrosinkinasaktivitet, och fosforylerar proteinet IRS-1 (bland annat).
7. Ett protein kallat GRB2 har SH2-domäner som binder till de numera fosforylerade residualerna på IRS-1. GRB2 binder också via SH3-domäner till ett GNRP (Gnukleotide releasing protein) vid namn Sos. På så sätt formeras ett aktivt komplex IRS1-GRB2-Sos.
8. Det aktiva Sos katalyserar ett G-proteins utbyte av GDP mot GTP.
9. G-proteinet aktiverar ett kinas kallat Raf, som fosforylerar ett MAPK-kinas (MAPKK). I och med det är vi inne i den signalväg som kallas MAPK-kaskaden.
10. MAPK-kinaset fosforylerar MAPK som för signalen vidare och i slutändan orsakar olika omställningar i cellens tillstånd, bl.a. med genreglering.

Se också Hancock, "Cell signaling", kap 9.

Lac-operonet

Ett exempel på relativt enkel genreglering finns i Lac-operonet i E.Coli, som bör vara bekant från tidigare kurser. Regleringen går ut på att minska uttrycket av huvudprodukten β -galaktosidas (som spjälkar laktos till glukos) då laktos inte finns tillgängligt eller glukos redan finns i relativt höga koncentrationer. Regleringen kan sammanfattas som:

1. Genen Lac-i uttrycker en repressor, som i sin nativa form binder till operonets operatorlocus (Lac-o) och därmed inhiberar transkription av operonets strukturgener z, a, och y.
2. Närvaro av β -galaktosider (laktos och diverse analoger) binder till repressorproteinet och hindrar därmed dess inhiberande verkan, och aktiverar därmed uttrycket av strukturproteinerna.
3. Närvaro av glukos orsakar en förhöjning i koncentrationen av intracellulärt cAMP (via diverse mekanismer som inte bryr oss om här). cAMP binder till och orsakar en konformationsförändring i repressorn CAP, som därmed får förmåga att binda till operatorlocus och inhibera genuttryck.

Se också Zubay, s.887-890.

Exempel på vanliga frågor

Som en guide för designen, så kan det vara bra att anpassa ert system så att det klarar följande vanliga frågor:

- Givet ett ämne X:

- Vilka signaler (reaktioner) förstärks/försvagas av ämnet?
- Givet en signal X:
 - Vilka signaler orsakar/orsakas av X?
 - Vilka signaler orsakar/orsakas av dessa i sin tur? (Obs. att detta är en rekursiv sökning i databasen. Se Elmasri & Navathe: Fundamentals of Database Systems, kap.22.1.3 (Ed5) om SQL3. För denna uppgift får ni anta att signalvägar är 1, 2 eller 3 steg långa.)

Då ni har kopplat ihop er databas med de övriga delprojekten bör det också gå att ställa mer övergripande frågor som exempelvis.

- Vilka enzymer deltar i en signalväg X?
- Vilka proteiner utgör dessa enzymer, och vilka gener kodar för dessa proteiner?

Samt ta fram en massa information om dessa enzymer och gener, förstås. Men det sköter de andra grupperna.

Referenser

Zubay, Geoffrey L: "Biochemistry", 1998. The McGraw-Hill Companies, Inc. ISBN: 0697219003

Hancock, John T: "Cell Signaling", 1997. Addison Wesley Longman Ltd. ISBN 0582312671