

# FetchProt

Kunskap om proteiner och deras funktioner är en grundförutsättning för, bland annat, all läkemedelsframställning. Information om proteiners funktioner finns inte samlad och strukturerad, utan är spridd framför allt över tusentals vetenskapliga artiklar. Inom projektet FetchProt samlas expertkunskap från områdena *bioinformatik*, *biokemi*, *datorlingvistik* och *informationsinfrastruktur* för att bygga ett system som automatiskt ska finna och sammanställa information om proteiners funktioner och de experiment som verifierat dessa, så att denna kunskap blir lättillgänglig för de forskare som behöver den.

Idag känner forskarsamfundet med säkerhet till **funktionen** hos endast en delmängd av de proteiner man är bekant med. I andra fall förutsägs ofta proteinets funktion genom att man tittar på strukturella likheter med proteiner vars funktioner är kända. Detta görs inte sällan i flera led, så att även proteiner vars funktion är endast förmodad, i sin tur används som jämförelseobjekt och på så vis kan avståndet i jämförelsekedjan bli långt till ett protein med **experimentellt verifierad** funktion.

Inom FetchProt kommer vi att använda följande ontologier för att göra kunskapsrymden hanterlig:

1. **Gene Ontology (GO)**. Målet med GO är att skapa en styrd terminologi för gener och genprodukter, som kan appliceras på alla organismer, oavsett att kunskapen om geners och proteiner roller i cellen ständigt utökas och förändras. GO tillhandahåller tre strukturerade nät av definierad terminologi som kan användas för att beskriva egenskaper hos gener och proteiner. Vi kommer i första hand att använda oss av det nät som beskriver proteiners molekylära funktion.
2. **En ontologi över experiment** som används för att verifiera proteinfunktion. Vi skapar en unik ontologi över olika typer av experiment som används idag för att undersöka proteinfunktioner.

Dessa två ontologier kopplas till varandra med en kunskapsrepresentation i form av regler som anger vilka kombinationer som är möjliga och tillräckliga för att ett protein ska anses ha en viss molekylär funktion.

De första stegen i FetchProt-systemets analys av varje dokument kommer att innefatta i tur och ordning:

1. *att identifiera textavsnitt som beskriver experiment i vetenskapliga artiklar*. Vi kommer att utveckla och utvärdera olika metoder för att automatiskt finna de deltexter i fulltextartiklar som beskriver enskilda experiment
2. *att identifiera vilket experiment det rör sig om, och om resultatet varit positivt*. Vi kommer att matcha syntaktiskt och morfologiskt analyserade representationer av dessa texter mot lingvistiskt baserade mönster som representerar de olika experiment som används för att testa proteiners funktion. Vi måste också kunna avgöra om experimentets resultat varit positivt
3. *att tilldela proteiner funktion efter vilka experiment som utförts med positivt resultat*. Regler appliceras som avgör om de positiva experiment som utförts kan koppla det omnämnda proteinet till någon viss molekylär funktion i GO