

# BILAGA 1: PROJEKTBEKRIVNING — FETCHPROT

## 1 Redogörelse för projektets syfte, mål och frågeställningar

### 1.1 Sammanfattning

Projektet *FetchProt* syftar till att förse det vetenskapliga samfundet, bioteknik- och läkemedelsindustrin med en allmänt tillgänglig infrastruktur för hantering och insamling av kunskap om proteiners funktioner.

Målet är att, genom att applicera språkteknologiska metoder för textanalys, automatisera processen att finna, värdera och samla in information om de proteiner som har en experimentellt verifierad funktion, från vetenskapliga artiklar inom molekylärbiologi och biokemi, samt att bygga upp infrastruktur och kunskapsbaser som gör denna information lättillgänglig och administrerbar.

### 1.2 Vi vet inte mycket om proteiners funktioner

Forskning om biologiska system förutsätter idag kunskap om arvsmassan. När man har bestämt arvsmassan, d.v.s. sekvenserat ett genom, består arbetet till stor del av att analysera DNA-sekvensen. Sedan metodiken för att sekvensera DNA blivit mer tillgänglig, är nu analysen av DNA-sekvensen den begränsande faktorn när man ska utvinna information från arvsmassan. Första steget är att identifiera gener i arvsmassan, var de befinner sig och hur långa de är. Generna översätts sedan till de proteiner som de kodar för och proteinernas funktioner analyseras.

Ett nyupptäckt proteins funktion fastställs ofta genom att titta på likheter med ett känt protein och dess redan fastställda funktion. Ibland är denna funktion experimentellt verifierad, men ofta nog är även detta proteins funktion bestämd genom likhet till ytterligare ett annat protein. Denna kedja kan bli lång och likheten mellan de proteiner som skall analyseras och de proteiner vars funktion är experimentiellt verifierade blir ibland allt mindre, vilket naturligtvis kan orsaka problem.

### 1.3 En ansats har gjorts för att samla informationen

Detta problem har rönt uppmärksamhet och 2001 publicerades databasen EXProt (Ursing et al., 2001) som endast innehåller proteiner med experimentellt verifierad funktion. Utgångspunkten var att samla proteinsekvenser med information om experimentell verifiering från olika befintliga databaser. Av alla organismers proteiner har man verifierat funktionen experimentellt hos omkring 20 000. Som en jämförelse kan nämnas att bara det mänskliga genomet kodar för uppskattningsvis 30 000 proteiner. EXProt innehåller för närvarande c:a 6 000 proteiner. Antalet proteiner i EXProt är alltså än så länge relativt lågt och begränsas av bristen på databaser som innehåller sådan information. Sedan projektet startades har dock stora genomikcentra börjat att med mycket resurser manuellt gå igenom alla proteiner i de större genomprojekten. Bland annat indikerar de för varje protein vilken grad av evidens som finns för den beskrivna funktionen.

### 1.4 Kunskap om verifierade proteinfunktioner finns dold i text

Då manuella litteraturstudier används för funktionsanalysen är denna tidskrävande och viss variation förekommer i klassificeringen mellan olika kuratorer. Med den ständigt ökande mängden publikationer är det dessutom svårt att hålla informationen uppdaterad. Mer än 2000 referenser

till vetenskapliga tidskriftsartiklar läggs till databasen MEDLINE<sup>1</sup> varje dag och antalet poster i databasen, nu cirka 11 miljoner, förväntas fördubblas varje år. Många av dessa artiklarna finns att tillgå i maskinläsbara format.

Syftet med EXProt är att samla information som nu finns spridd i ett flertal, heterogena informationskällor tillgängliga över elektroniska nätverk, på ett enda ställe. Denna resurs skall kunna användas för att på ett säkrare sätt kunna förutsäga funktionen hos proteiner med okänd funktion, och kommer därigenom att ge ökade möjligheter att förstå biologiska samband.

Målet med detta projekt är att ur elektroniskt publicerade vetenskapliga artiklar med hjälp av datoriserad textanalys automatiskt identifiera proteiner med en experimentellt verifierad funktion samt att samla in sekvensdata för dessa proteiner. Det framtagna systemet blir en viktig infrastrukturell resurs som kommer att bli mycket användbar vid analysen av okända proteiner och möjliggöra en mer komplett bild av tillgänglig information, samt leda till att denna enklare kan hållas uppdaterad och spridas.

## 1.5 Projektets relation till forskningsfronten

Under senare år har det växt fram ett stort forskningsfält i skärningspunkten mellan bioinformatik och datorlingvistik, ibland kallat bioNLP. Man har insett att eftersom de överblickbara mängder kunskap som produceras inom livsvetenskaperna inte främst finns i strukturerade databaser utan i hög grad i de vetenskapliga artiklar som publiceras inom området, så krävs språkbaserade verktyg för att gå in i texten och skörda den information som finns där. Detta speglas i det ökade intresset för datorlingvistiska metoder inom det biomedicinska forskningssamfundet och för den biomedicinska domänen bland datorlingvisterna. Många av de större konferenserna inom livsvetenskaperna har infört speciella teman om språkteknologiska metoder och likaså har närapå alla språkteknologiska konferenser de senaste två åren haft spår för datorlingvistisk forskning och språkteknologiska tillämpningar inom den biomedicinska domänen. De språkteknologiska forskningsansatserna har i mycket fokuserat på två områden: identifiering av textomnämnanden av biomedicinska entiteter och identifiering av relationer entiteterna emellan. En specifik texttyp dominerar som undersökningsobjekt: sammanfattningar av vetenskapliga artiklar från ovannämnda databas, MEDLINE. Det aktuella projektet siktar på flera sätt högre än dessa mål, dels genom att söka efter mer komplexa företeelser (uttryck för experimentellt verifierade proteinfunktioner) i texten, som därför måste generaliseras på en högre abstraktionsnivå, dels genom att bearbeta hela vetenskapliga artiklar snarare än sammanfattningar av artiklar.

Språkteknologigruppen på Swedish Institute of Computer Science (SICS) har genom ett av VINNOVA delvis finansierat projekt, *Proteinhalt i text*<sup>2</sup>, fått en god överblick över detta område och har presenterat resultat från projektet på en rad av dessa bioNLP-konferenser (Franzén et al., 2002; Eriksson et al., 2002; Lidén et al., 2002; Olsson et al., 2002). Inom detta projekt utvecklades ett verktyg – Yapex – som automatiskt identifierar och märker upp proteinnamn i biomedicinsk text. De tekniker som användes i det projektet kommer att ligga till grund för de algoritmer som kommer att utvecklas inom det aktuella projektet.

Problemet som ska lösas inom detta projekt kan relateras till de problem som forskningsområdet Informationextraktion (IE) ägnar sig åt. IE är ett språkteknologiskt forskningsområde som började definieras i slutet av 80-talet genom en serie konferenser (MUCs — Message Understanding Conferences) som löpte fram till 1997 (Sundheim, 1991; Sundheim, 1992; Sundheim, 1993; Sundheim, 1995; Chinchor, 1998). Uppgiften för ett IE-system är att ur löpande text extrahera så mycket information som möjligt om en viss förutbestämmd typ av händelse, för att sedan strukturera informationen i ett entydigt format för vidare behandling. Exempel på händelser som det har byggts IE-system för är terroristattacker, flygplansolyckor, raketuppskjutningar samt förändringar på högre positioner inom näringslivet. Det är uppenbart att extraktion av "händelsen" *experimentell verifiering av proteinfunktioner* också kan ses som en IE-uppgift som kan lösas med IE-metoder. Språkteknologigruppen på SICS har gedigen erfarenhet av IE-system (Franzén, 1999; Franzén et al., 2002) och har arbetat

<sup>1</sup>MEDLINE är en bibliografisk databas med medicinska vetenskapliga artiklar sökbar via PUBMED: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>

<sup>2</sup>Se <http://www.sics.se/humle/projekt/prothalt/>

i nära samarbete med flera av de viktigaste forskargrupperna inom området. SICS driver också internt ett pågående arbete med en öppen och generell arkitektur för informationsförädling kallad Kaba (Olsson, 2002) som kommer att vidare förfinas under detta projekt (se avsnitt 4.2.1).

Databasen EXProt är unik såtillvida att den är det enda försöket att samla denna typ av information på ett ställe.

## 2 Bedömning av projektets relevans

### 2.1 Övergripande relevans

Det föreliggande projektförslaget har stor relevans för Sveriges vidare profilering inom bioinformatik från ett språkteknologiskt perspektiv — i omvärlden satsas det på bioNLP-området och resultaten från FetchProt kan, som fortsättning på det framgångsrika projektet *Proteinhold i text*, ytterligare synliggöra svensk språkteknologi inom ett framväxande och intressant forskningsområde.

Resultaten av projektet kommer att kunna generaliseras till att gälla andra domäner än den projektspecifika och därigenom visa på nyttan av de systemlösningar som utvecklas i projektet och på användbarheten av språkteknologi i kunskapshanteringssystem.

### 2.2 Relevans i relation till specifika intressenter inom näringsliv, arbetsliv och samhälle

Då resultaten från projektet i så stor utsträckning som möjligt kommer att göras fritt tillgängliga, kommer de delar av svenskt näringsliv, arbetsliv och samhälle som har intressen inom den bioinformatiska domänen att kunna tillgodogöra sig dessa. I Sverige finns aktörer som är stora på den internationella marknaden för biomedicin och som i sin dagliga verksamhet arbetar med bioinformatiska frågor; vi antar, på god grund, att dessa aktörer har nytta av sätt att effektivisera och underlätta hanteringen av dessa frågor, något som resultaten från FetchProt kan hjälpa dem med. Likaså kommer den framgångsrika svenska forskningen inom det biomedicinska området att kunna dra nytta av resultaten från projektet.

## 3 Förväntade resultat och effekter på kort och lång sikt

### 3.1 Effekter på kort sikt (under projektets gång)

- Kunskap om språkteknologiska tekniker och metoder både generellt och specifikt för bioinformatik, kommer att öka bland intresserade organisationer tack vare öppna seminarier och rapporter som avhandlar projektets resultat.
- Fria infrastrukturella språkteknologiska resurser kommer att göras tillgängliga i form av data och implementerade algoritmer.
- Projektets tvärvetenskapliga karaktär kommer att leda till fruktbart kunskapsutbyte mellan projektparterna.

### 3.2 Effekter på lång sikt (i samband med projektslut och därefter)

- En mer omfattande version av databasen EXProt kommer att göras tillgänglig för forskare och näringsliv, vilket kommer att resultera i bättre förutsättningar för molekylärbiologisk och biokemisk forskning.
- En ökande kunskapsmassa kring bioNLP i Sverige kommer att leda till möjligheter för företag att utveckla nya tjänster att erbjuda sina kunder; resultaten från projektet kan komma att exploateras i kommersiella produkter.

- Resultaten från projektet kan ge implikationer för språkvetenskapliga teorier om den biomedicinska domänens beskaffenhet som texttyp.
- Erfarenheterna från byggandet av ett distribuerat informationssystem kommer att stärka konkurrenskraften hos den medverkande företagspartnern.

## 4 Projektplan

Projektet är tänkt att löpa över tre år. Projektet har tre parter med tre olika huvudkompetenser:

- Swedish Institute of Computer Science (SICS) är huvudsökande och ansvarig för utveckling av de språkteknologiska analysverktygen samt för samordning av projektet.
- Centrum för genomik och bioinformatik (CGB) vid Karolinska Institutet är ansvarig för databasen EXProt, för att tillhandahålla textmaterial och för domänkunskapen.
- Metamatrix (MMX) är ansvariga för systemarkitektur, systemdesign samt systemimplementation.

### 4.1 Angreppssätt

Projektet kommer att inledas med att fastställa definitionen av *experimentellt verifierad funktion*. Denna definition liksom en grundlig, datadriven, empirisk studie av hur experimentell evidens för proteiners funktion realiserar i text kommer att utgöra basen för projektet. Det är därför av yttersta vikt att dessa två moment inte förfuskas. Vi vill betona betydelsen av ett nära samarbete mellan SICS och CGB under hela projektförloppet för att garantera att den domänkunskap som CGB besitter tas tillvara och implementeras korrekt i de domänspecifika språkmodulerna. För att garantera projektets empiriska grund kommer ett uppmärkningsarbete av text att genomföras. Därigenom får vi ett facit, uppdelat i en referenskorpus och en utvärderingskorpus, mot vilket vi kan träna och utvärdera extraktionsmodulerna i systemet. Utvecklingen av textanalyskomponenterna kommer att genomföras i en iterativ process där de enskilda modulerna utvärderas och förbättras kontinuerligt under projektets gång.

På ett tidigt stadium är det också viktigt att sammanställa en kravspecifikation för den systemlösning som ska utvecklas inom projektet. Det är mycket angeläget att MMX grundar systemets arkitektur och design på principer som överensstämmer med de professionella användarnas förväntningar och domänens konventioner. För att garantera skalbarhet, långsiktig hållbarhet och flexibilitet är det viktigt att systemet bygger på väldefinierade komponenter vilka kopplas samman med hjälp av standardiserade kommunikationsprotokoll. Systemet måste vidare kunna hämta information från en heterogen samling informationskällor. Gränssnitten måste upplevas som enkla och naturliga av användarna.

Vi kommer att vinnlägga oss om att göra resultaten, både vad gäller textanalys och systemarkitektur, så generaliserbara att de kan nå bredare användning utanför projektets specifika fall.

### 4.2 Metoder och verktyg

Väl beprövade metoder från IE-området kommer huvudsakligen att användas i textanalysmodulerna. Proteinnamnsidentifieraren Yapex kommer att vara en viktig komponent, liksom de verktyg för syntaktisk och morfologisk analys som vi planerar att använda. Dessa komponenters utdata kommer att bedömas och bearbetas på olika sätt i moduler uppbyggda med hjälp av Kaba.

#### 4.2.1 Kaba

Det flesta av textanalysuppgifterna som ska utföras i projektet kommer att genomföras med hjälp av Kaba, en öppen och generell arkitektur för informationsförädling som utvecklas av SICS informationsåtkomsttema.

Kaba är en öppen och modulär arkitektur för generell textanalys och informationsförädling, som började utvecklas i avsikt att bygga upp ett informationsextraktionssystem. Efter hand har det visat sig att systemet kan generaliseras till att bli användbart för många andra typer av textanalys och informationsförädling. Kravspecifikationen och designen av Kaba beskrivs av Olsson (2002).

Kaba bygger på resultat från TIPSTER-gruppens arbete om hur dokument bör kunna representeras och bearbetas (Grishman et al., 1997). Kaba implementeras helt och hållet i Java för att göra det plattformsoberoende och kärnan i systemet ska distribueras gratis under en mycket fri licens.

Kaba är inte byggt för att göra lingvistisk analys på låg nivå, som till exempel syntaktisk och morfologisk analys, utan snarare för att analysera och bearbeta text efter att en sådan analys har gjorts.

#### 4.2.2 Funktionell dependensgrammatisk analys

Grundläggande språklig analys i projektet görs av moderna hjälpmedel för syntaktisk och morfologisk analys utvecklade hos Connexor Oy, SICS mångåriga samarbetsparter i Helsingfors. Den grundläggande analysen baserar sig på dependensgrammatik och lämpar sig mycket väl till fortsatt behandling för extraktion av entiteter och deras bestämningar i text; även delanalyser är användbara i de fall inte hela satser kan analyseras.

### 4.3 Tids- och arbetsplan

Tidsplanen i Tabell 1 visar varje arbetspakets (AP) utsträckning månadsvis och omfattar totalt knappt ett och ett halvt personår per år i tre år. Ett arbetspakets utsträckning i tiden innebär inte nödvändigtvis att aktivitetsgraden i det är konstant, den varierar beroende på interaktionen mellan olika arbetspaket. I beskrivningen av respektive arbetspaket nedan anges dess arbetsomfattning i personmånader (pm).

Projektplanen gäller givet att ansökan beviljas enligt budgetförslaget (se avsnitt 7).

**AP 1: Agent för textinsamling.** Verktyg för kontinuerligt tillflöde av nytt textmaterial till analysprocessen. För att uppfylla kravet på kontinuerlig uppdatering av EXProt-databasen krävs ett högpresterade verktyg i form av en autonom agent för återkommande insamling av information från olika typer av informationskällor (databaser, webbsidor etc.). Arbetet kommer att pågå cykliskt under projekttiden med tonvikt på den inledande fasen. Detta AP innefattar också analys och beslut om vilka källor som texterna ska hämtas från. **Beroende:** inget. **Deltagare:** MMX, CGB. **Omfattning:** 3 pm. **Resultat:** implementerad insamlingsagent, textdata.

**AP 2: Komponent för textförbearbetning.** Implementering av en komponent som bearbetar den insamlade informationen under AP1 så att den görs tillgänglig för textanalysverktyget, (AP5). **Beroende:** AP1. **Deltagare:** MMX. **Omfattning:** 0,5 pm. **Resultat:** implementerad textförprocessor.

**AP 3: Analys av målbegrepp.** Definition av målbegreppet *experimentellt verifierad proteinfunktion*. Arbetet i detta paket innebär att fastställa vad det innebär för en proteinfunktion att vara experimentellt verifierad och att analysera och sammanställa alla de sätt på vilka detta kan uttryckas i vetenskaplig text. Arbetet kommer att baseras på litteraturstudier och på analys under annoteringsarbetet av projekttextkorpusen och är en förutsättning för analys- och annoteringsarbetet under AP4, men kommer samtidigt också att påverkas av detsamma. **Beroende:** AP4. **Deltagare:** SICS, CGB. **Omfattning:** 4,5 pm. **Resultat:** kravspecifikation, rapport.

**AP 4: Datainsamling och bearbetning.** Sammanställning, analys och annotering av referens- och utvärderingskorpus. På grundval av arbetet under AP3 sammanställs och analyseras en textkorpus bestående av textmaterial av samma typ som kommer att behandlas av resultatet från AP2. Korpusen kommer att sammanställas enligt projektets behov och kommer efter annotering utgöra grund för utvecklings- och utvärderingsarbetet under AP5 och AP6.

AP 11	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
AP 9											⊗	⊗
AP 8	⊗	⊗				⊗	⊗	⊗	⊗	⊗	⊗	⊗
AP 7											⊗	⊗
AP 6									⊗	⊗		
AP 5				⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
AP 4	⊗	⊗	⊗	⊗	⊗							
AP 3	⊗	⊗	⊗	⊗	⊗	⊗		⊗		⊗		
AP 2			⊗	⊗	⊗							
AP 1	⊗	⊗	⊗	⊗								
MÅNAD	1	2	3	4	5	6	7	8	9	10	11	12

AP 11	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
AP 9			⊗			⊗			⊗			⊗
AP 8	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
AP 7				⊗	⊗							
AP 6			⊗			⊗			⊗			⊗
AP 5	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
AP 3	⊗		⊗		⊗			⊗			⊗	
MÅNAD	13	14	15	16	17	18	19	20	21	22	23	24

AP 11	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
AP 10											⊗	⊗
AP 9			⊗			⊗						⊗
AP 8	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗				
AP 6			⊗			⊗			⊗	⊗		
AP 5	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗				
MÅNAD	25	26	27	28	29	30	31	32	33	34	35	36

Table 1: Arbetspaketens fördelning över tiden.

**Beroende:** AP3. **Deltagare:** SICS, CGB. **Omfattning:** 3,5 pm. **Resultat:** rapport, textkorpus.

**AP 5: Specifikation och implementation av textanalysverktyg.** Utgående från specifikationen från AP3 utformas en fullständig kravanalys och specifikation av funktionerna hos ett verktyg som uppfyller syftena med projektet (beskrivna i avsnitt 1). En första version av prototypen implementeras och driftsätts under år 1 och nya versioner utförs i en iterativ process baserade på resultat från utvärdering (AP6). **Beroende:** AP3, AP4, AP6. **Deltagare:** SICS, CGB. **Omfattning:** 13 pm. **Resultat:** programvara.

**AP 6: Kvantitativ och kvalitativ utvärdering.** Kvantitativ utvärdering av textanalysverktyget gentemot resultatet av datainsamlingen utförd i AP4. Kvalitativ utvärdering av textanalysen med avseende på tillförlitlighet beträffande verifikationsbedömning. Under senare delar av projektet kommer systemarkitektur och integration, gränssnitt och samarbete mellan systemets komponenter att utvärderas. **Beroende:** AP1, AP2, AP3, AP4, AP5, AP7. **Deltagare:** SICS, CGB, MMX. **Omfattning:** 5 pm. **Resultat:** rapporter.

**AP 7: Gränssnitt mellan textanalys och databas.** Implementering av en komponent som bearbetar resultatet från textanalyskomponenten, (AP5), och hanterar informationen om verifierad proteinfunktion som ska tillföras databasen. **Beroende:** AP5. **Deltagare:** MMX, CGB. **Omfattning:** 3 pm. **Resultat:** programkomponent.

**AP 8: Systemarkitektur, integration och användargränssnitt.** Arbetspaketet syftar till att designa och implementera ett distribuerat system, i möjligaste mån byggt på öppna industristandarder såsom SOAP/XML-RPC och resulterar i ett publicerat system med webbaser-

ade användar- och administratörsgränssnitt, systemdokumentation med API-definition och webbtjänstspecifikation. **Beroende:** AP1, AP2, AP5, AP7. **Deltagare:** MMX, SICS, CGB. **Omfattning:** 10 pm. **Resultat:** Publicerat system.

**AP 9: Periodisk rapportering.** Fortlöpande rapportering och av projektresultat genom SICS seminarieserier öppna för avnämare och aktuella och tänkta intressenter, och genom periodisk vetenskaplig publicering av resultat. **Beroende:** samtliga arbetspaket. **Deltagare:** SICS, CGB. **Omfattning:** 4 pm. **Resultat:** serie öppna seminarier, serie internationella vetenskapliga publikationer.

**AP 10: Slutrapportering.** Avrapportering och summering av projektet till VINNOVA och övriga intressenter. Kunskapsspridning genom ett eller flera seminarier, offentliga forskningsrapporter och tillgängliggörande av programvara till allmänheten. **Beroende:** samtliga arbetspaket. **Deltagare:** SICS, CGB, MMX. **Omfattning:** 2,5 pm. **Resultat:** öppet slutseminarium, skriftlig slutrapport, allmänt tillgänglig programvara.

**AP 11: Projektadministration.** Tid avsatt för administration och koordination av arbetsinsatser, regelbundna projektmöten, sammanställning av interna delrapporter och kontakter med intressenter. **Beroende:** inget. **Deltagare:** SICS. **Omfattning:** 3 pm.

## 5 Deltagande personers kompetens och erfarenhet av uppgiften

### 5.1 SICS

Swedish Institute of Computer Science, SICS, är ett icke-vinstdrivande forskningsinstitut. SICS bidrar till konkurrenskraften hos svensk industri genom att bedriva avancerad datavetenskaplig forskning inom valda områden och att aktivt föra ut resultaten till industrin. På SICS arbetar ca 100 forskare som förenar djupt forskningsintresse med en vilja att se forskningsresultaten omsättas i praktiskt fungerande lösningar, produkter och tjänster. Vi forskar därför i nära samarbete med både de stora, internationellt ledande företagen och små, ofta nystartade företag.

Personer på SICS som är aktuella för det här projektet presenteras härnedan. Det är dock inte bestämt exakt hur projektet kommer att bemannas, varför det är svårt att redogöra för varje persons arbetsinsats.

#### 5.1.1 Jussi Karlgren

Jussi Karlgren har doktorerat i datorlingvistik vid Stockholms universitet, och har arbetat som forskare på SICS sedan 1990, främst med frågor kring informationsåtkomst och dialoger mellan människa och maskin. Han har tidigare arbetat på SISU, varit forskarstuderande vid Columbia University, arbetat som gästforskare på Xerox PARC och New York University och varit t f professor i språkteknologi vid Helsingfors universitet, institutionen för lingvistik. Jussi har också varit koordinator för det inom EUs femte forskningsprogram finansierade projektet LLAVES.

#### 5.1.2 Kristofer Franzén

Kristofer Franzén är doktorand i datorlingvistik vid Stockholms universitet, institutionen för lingvistik. Under året 1997-1998 arbetade han som gästforskare vid New York University, datavetenskapliga institutionen i projektet Proteus där han anpassade ett befintligt informationsextraktionssystem till svenska. Han är sedan 1998 forskare på SICS och har arbetat främst med allmän informationsextraktion, Kaba-systemet och språkteknologiska frågor i den biomedicinska domänen.

### 5.1.3 Gunnar Eriksson

Gunnar Eriksson är doktorand i allmän språkvetenskap vid Stockholms universitet, institutionen för lingvistik, där han under åren 1988-1999 arbetade i ett antal datorlingvistiska och allmänlingvistiska forskningsprojekt. Åren 2000-2001 arbetade Gunnar vid Nordisk Språkteknologi AS med lingvistiska aspekter på taligenkänning. Sedan 2001 är han forskare vid SICS där han främst arbetat med informationsextraktion inom den biomedicinska domänen: utvecklandet av proteinnamnsigenkännaren Yapex.

## 5.2 Centrum för genomik och bioinformatik

### 5.2.1 Björn Ursing

Björn Ursing är doktor i genetik och har postdoc-erfarenhet av bioinformatik. Björn är gruppleddare på *Centrum för genomik och bioinformatik* vid Karolinska Institutet med sekvensanalys och genomannotering som specialitet. Han arbetade med annoteringen vid analysen av genomet av *Lactobacillus plantarum* under sin postdoc i Nederländerna. Det arbetet lade grunden för EXProt som var svaret på ett direkt behov vid annoteringsarbetet. Björn var initiativtagare till EXProt och har stått för en stor del av utvecklingen. Björn arbetar också deltid på Medivir AB som ansvarig för bioinformatiken.

## 5.3 Metamatrix

### 5.3.1 Patrik Jonasson

Patrik Jonasson, IT-bibliotekarie, systemanalytiker. Patrik har en mångårig erfarenhet som projektledare och kreatör inom området IT i skolan, såväl inom landet som i EU-projekt. Han har på senare tid alltmer fokuserat på projektledning, användningsfallsmodellering och utredningar. Patrik har medverkat i uppdrag hos bl.a. Nätuniversitetet, Högskoleverket, Valdemarsviks kommun och Stockholm Visitors Board. Patrik har haft samordningsansvar för Metamatrix forskningsprojekt och har därför goda kontakter med forskningsinstitut och forskningsfinansiärer.

### 5.3.2 Jonas Salling

Jonas Salling, högskolestudier i data- och informationsteknik vid KTH, systemarkitekt. Jonas har flerårig erfarenhet av objektorienterad design och programmering. Han har arbetat som lärare på KTH. Jonas har bred teknisk erfarenhet av olika språk, utvecklingsverktyg, applikationer, operativsystem etc. Jonas har medverkat i uppdrag hos bl.a. Skolverket, Högskoleverket, Nätuniversitetet och CFL.

### 5.3.3 Tobias Lidskog

Tobias Lidskog, civ.ing., systemutvecklare. Tobias har erfarenhet av olika operativsystem, utvecklingsmiljöer, programspråk och applikationer. Tobias har medverkat i uppdrag hos bl.a. Skolverket, Högskoleverket, Nätuniversitetet och Enköpings kommun.

## 6 Omfattningen av deltagarnas arbetsinsatser

- SICS andel i projektet uppgår till 55% av en heltid under 3 år. 5% av detta är budgeterat för löneklassen EXPERT och 50% för löneklassen SENIOR.
- CGB:s andel i projektet uppgår till 50% av en heltid under 3 år. 5% av detta är budgeterat för löneklassen FORSKARE och 45% för motsvarande löneklassen DOKTORAND.
- MMX andel i projektet uppgår till 1 800 timmar fördelat över de tre åren.

## 7 Totalbudget för projektet

En detaljerad budget för projektet presenteras i Tabell 2.

<b>Kostnader kkr</b>					
	<b>2003</b>	<b>2004</b>	<b>2005</b>	<b>2006</b>	<b>Summa</b>
SICS löner	113	338	338	225	<b>1014</b>
SICS utrustning	0	0	0	0	<b>0</b>
SICS material, drift	0	0	0	0	<b>0</b>
SICS resor	10	30	30	10	<b>80</b>
SICS övrigt (se bilaga 3)	104	313	313	209	<b>939</b>
<b>Summa SICS</b>	<b>227</b>	<b>681</b>	<b>681</b>	<b>444</b>	<b>2033</b>
<b>Externa tjänster CGB</b>					
Löner	54	162	162	108	<b>486</b>
Resor	10	30	30	10	<b>80</b>
Utrustning	85	20	10	5	<b>120</b>
Övrigt	10	20	20	10	<b>60</b>
Förvaltning, lokalhyra	56	81	78	47	<b>262</b>
<b>Summa CGB</b>	<b>215</b>	<b>313</b>	<b>300</b>	<b>180</b>	<b>1008</b>
<b>Externa tjänster MMX</b>					
Löner	80	240	240	160	<b>720</b>
<b>Summa MMX</b>	<b>80</b>	<b>240</b>	<b>240</b>	<b>160</b>	<b>720</b>
<b>Summa Kostnader</b>	<b>522</b>	<b>1234</b>	<b>1221</b>	<b>784</b>	<b>3761</b>
<b>Finansiering kkr</b>					
Insats CGB	10	30	30	20	<b>90</b>
Insats MMX	40	120	120	80	<b>360</b>
Anslag VINNOVA	472	1084	1071	684	<b>3311</b>
<b>Summa Finansiering</b>	<b>522</b>	<b>1234</b>	<b>1221</b>	<b>784</b>	<b>3761</b>

Table 2: Sammanställning kostnader och finansiering.

**Observera** att lönekostnader och naturinsatser i form av arbetstid för Metamatrix AB (MMX) i Tabell 2 är baserade på en schablonkostnad uppgående till 600 kr/h.

**Observera även** att de indirekta kostnaderna under "Övrigt" täcker SICS kostnader för ledning, administration, post, dator- och telekommunikation, bibliotek, lokaler, samt delade datorresurser

## 8 Kommunikationsinsatser

För att sprida de teoretiska resultaten från projektet planerar vi att hålla öppna seminarier, samt att publicera oss i olika vetenskapliga sammanhang. Som ett led i kunskapsspridningen kommer vi också att göra informationsförädlingsarkitekturen Kaba fritt tillgänglig för allmänheten, samt i separata moduler de språkteknologiska delresultat — lexika, tesauri, extraktionsalgoritmer — vi utvecklat under projektets gång. Projektets utveckling och resultat kommer att kunna följas på projektets hemsida.

## 9 Miljökonsekvenser

Det föreslagna projektet kan inte förutses ge direkta konsekvenser för inre eller yttre miljö.

## Referenser

- Chinchor, Nancy A. editor. 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Virginia, USA, April-May. Morgan Kaufmann. [http://www.itl.nist.gov/iaui/894.02/related\\_projects/muc/proceedings/muc\\_7.toc.html](http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7.toc.html).
- Eriksson, Gunnar; Franzén, Kristofer; Olsson, Fredrik; Asker, Lars och Lidén, Per. 2002. Using heuristics, syntax and a local dynamic dictionary for protein name tagging. In *Proceedings of Human Language Technology 2002*, San Diego, USA, March 24-27.
- Franzén, Kristofer; Eriksson, Gunnar; Olsson, Fredrik; Asker, Lars; Lidén, Per och Cöster, Joakim. 2002. Protein names and how to find them. *International Journal of Medical Informatics*, 67(1-3):49-61.
- Franzén, Kristofer. 1999. Adapting an English information extraction system to Swedish. In *Proceedings of the 12th Nordic Conference of Computational Linguistics*, pages 57-65, Norwegian University of Science and Technology, Trondheim, Norway, December.
- Franzén, Kristofer; Eriksson, Gunnar; Olsson, Fredrik; Asker, Lars och Lidén, Per. 2002. Exploiting syntax when detecting protein names in text. In *Proceedings of Workshop on Natural Language Processing in Biomedical Applications*, Nicosia, Cyprus, March 8-9.
- Grishman, Ralph; Dunning, Ted; Callan, Jamie; Caid, Bill; Cowie, Jim; Guthrie, Louise; Hobbs, Jerry; Jacobs, Paul; Mettler, Matt; Ogden, Bill; Schwartz, Bev; Sider, Ira och Weischedel, Ralph, 1997. *TIPSTER Text Phase II Architecture Design. Version 2.3*. New York, New York, January.
- Lidén, Per; Asker, Lars; Eriksson, Gunnar; Franzén, Kristofer och Olsson, Fredrik. 2002. Protein name tagging for browsing support, active database cross linking and information retrieval. In *Proceedings of Bioinformatics 2002*, Bergen, Norge, April.
- Olsson, Fredrik; Eriksson, Gunnar; Franzén, Kristofer; Asker, Lars och Lidén, Per. 2002. Notions of correctness when evaluating protein name taggers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, 24 August - 1 September.
- Olsson, Fredrik. 2002. *Requirements and Design Considerations for an Open and General Architecture for Information Refinement*. Licentiate of philosophy thesis, Department of Linguistics, Uppsala University, Uppsala, March.
- Sundheim, Beth editor. 1991. *Proceedings of the Third Message Understanding Conference (MUC-3)*. Morgan Kaufman, May.
- Sundheim, Beth editor. 1992. *Proceedings of the Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufman, June.
- Sundheim, Beth editor. 1993. *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, Baltimore, Maryland, USA, August. Morgan Kaufman.
- Sundheim, Beth editor. 1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, Maryland, USA, November. Morgan Kaufman.
- Ursing, Björn M.; Enckevort, van Frank H. J.; Leunissen, Jack A. M. och Siezen, Roland J. 2001. EXProt — a database for EXperimentally verified Protein functions. *In Silico Biol.*, 2(0001). <http://www.bioinfo.de/isb/2001/02/0001/>.