Interessetilkendegivelse: ” National Life Science Supercomputer”

Den danske ESFRI ELIXIR node - Infrastruktur for biologiske data, analyse og integration

Indledning. De biologiske, bioteknologiske og medicinske forskningsområder har i det sidste 10 år skiftet karakter, idet der nu i langt højere grad end tidligere findes data hvis omfang gør det både muligt og nødvendigt at tage ny informationsteknologi i brug. I stort set alle områder af den biologiske og medicinske forskning er computerens rolle blevet mere og mere markant. Først og fremmest har de store genomprojekter, hvor arveanlæg er blevet bestemt for mange forskellige organismer, på kort tid gjort behovet for kompetence indenfor bioinformatisk supercomputing meget større.

I relation til det humane genom skifter fokus for tiden til den humane variation, hvor perspektivet ikke blot er at opdage forhold, der er generelle for alle mennesker, men netop det der er specifikt for det enkelte individ. Selvom det oprindeligt tog mere end 10 år at bestemme en ”gennemsnitsudgave” af de menneskelige arveanlæg (til en udgift i omegnen af 20 milliarder kroner), er prisen snart nogle få tusinde kroner. Derved opstår der nye muligheder for at undersøge hvordan det enkelte individs særlige versioner af generne kan ændre forskellige processer, særligt i forbindelse med sygdomme og lægemidler, og dermed skabe viden der kan benyttes til design af individualiseret medicin. Målet er at bruge denne viden til at forbedre og forlænge menneskeliv ved at udnytte data og computere til det yderste.

Den første bølge af bioinformatisk computeranalyse fokuserede på basal sekvens- og strukturanalyse af enkelte gener, og inden for dette område vil der stadig være mange udfordringer, men først og fremmest vil mange af de fremtidige behov imidlertid dreje sig om sofistikeret integration af forskellige typer af data med høj diversitet, og den tilhørende "many task supercomputing". Interessen for det enkelte makromolekyle vil tydeligvis vare ved, men udvides til netværk af vekselvirkende makromolekyler og de dynamiske ændringer der sker i disse netværk som følge af ekstern påvirkning, eller som følge af genversioner med patogene egenskaber.

Den industrielle bioteknologi er også inde i en rivende udvikling, hvor sekventering af metagenomer indeholdende titusindvis af forskellige organismer har givet et gennembrud for alle de organismer, der ikke kan kultiveres i et laboratorium. Disse datamængder er enorme, og andre typer data, f.eks. metatranscriptomics data, der beskriver millioner af mikrobielle geners ekspression, gør disse beregningsopgaver meget komplekse fordi data er heterogene, og at der samtidigt skal mange forskellige algoritmer i spil i jagten på industrielle enzymer, nye typer cellefabrikker og indsigt i mikrobiologiens samspil med værtsorganismer. Mikrobiel genomsekventering er således inde i rivende udvikling, der giver nye bioteknologiske muligheder, der vil være essentielle i forhold til at etablere mere bæredygtige samfund, der ikke er baseret på olie og gas. Analysen af mikrobielle data med systembiologiske værktøjer er meget væsentlig, og vil have betydning for indsigt i økologiske nicher, som for eksempel tarmfloraen, og dens opførsel i sunde og syge individer.

Stor brugergruppe - demokraticering af supercomputing. EU og særligt DG Connect har i Horozon2020 prioriteret at supercomputing skal udbredes og gavne forskning og innovation mere bredt end det tidligere har været tilfældet. Her er de biologiske videnskaber en god case da supercomputingbehovet indenfor de

biologiske videnskaber har en anden brugerprofil når man sammenligner med mange andre forskningsområder. Sammenlignet med den gruppe af forskere der f.eks. arbejder med LHC data er der tale om en meget bredere brugerprofil, der involverer millioner af forskere bare i Europa. Omsat til danske forhold er der også tale om meget store miljøer som benytter supercomputerdrevne services, ligesom eksplosionen i lokal dansk generering af f.eks. DNA sekventeringsdata i størrelsesordenen over 100TB, har gjort at behovet for dedikerede workflows og kapacitet i Danmark er steget betydeligt. Helt nye forskningsområder har således fået behov for HPC. På baggrund af den massive stigning i antallet af Life Science brugere af HPC drevne resourcer (over 3 millioner brugere i Europa) og de internationale HPC centres brugerprofil (PRACE), der typisk tilgodeser erfarne brugere med optimerede koder indenfor mere homogene datadomæner, som man kender det fra fysikken og astronomien, motiverer vi oprettelsen af et nationalt Life Science supercomputing center.

Den danske ELIXIR node. Bioinformatikgrupperne på de danske universiter har sammen med ledende bioinformatikere i den danske biotek og pharma industri gået sammen om at opbygge en dansk node i den fælleseuropæiske infrastruktur for bioinformatik og biologisk information ELIXIR. ELIXIR har idag medlemskab af 16 lande, og det er sandsynligt at de fleste lande i løbet af få år vil tilslutte sig samarbejdet og ratificere formelt medlemskab. Infrastrukturen koordineres fra det Europæiske Bioinformatikinstitut Cambride, England.

Den danske nodes arbejde repræsenterer en vigtig forskningsinfrastruktur, der er væsentlig i den danske, nationale strategi på området, ikke blot på kort sigt, men også på mellemlang sigt (3-10 år). Det primære forskningsområde er Biotek, Sundhed og Life Science, dog med betydelige kontaktflader til Materiale og Nanoteknologi, Energi, Klima og Miljø, samt e-Science generelt. Etableringen af en bioinformatikinfrastruktur har været i en forberedende fase i et par år på europæisk niveau i ESFRI regi. Den den danske deltagelse i infrastrukturen er sikret i form af en infrastrukturbevilling på 25 M DKK til netop dette formål. Formålet med bioinformatikinfrastrukturen er at udbygge og sikre dansk deltagelse i ELIXIR, og at skabe og drive en stabil ramme for biologiske og medicinske databaser og relaterede softwareværktøjer som kan støtte biologisk forskning – både i relation til grundforskningen og til dens anvendelser indenfor medicin, bioteknologi og miljø, i den bioteknologiske industri og i samfundet som helhed. Både store, mellemstore og mindre danske virksomheder deltager i arbejdet. Generelt udgør det bioteknologiske og medicinske område en større og større del af det danske bruttonationalprodukt. Infrastrukturen er med i det danske infrastruktur roadmap, og der er brug for at sikre indsatsen, synligheden og stabiliteten på både mellemlang og lang sigt ved at tilføje indsatsen massive supercomputing resourcer.

De danske bidrag til den fælles europæiske forskningsinfrastruktur er fokuseret omkring værktøjer til analyse af data, herunder deres interoperabilitet og integration af data, men i en situation hvor Danmark, f.eks. med kinesiske og andre europæiske samarbejdspartnere i stigende grad producerer biologiske data, er databaseinfrastrukturen også meget væsentlig nationalt. Det er vigtigt, at Danmark sikres indflydelse, og at der sikres relevante danske forskningsmiljøer adgang til dedikeret supercomputing kapacitet - hele området har afgørende interesse og betydning for Danmark økonomi både på kort sigt og langt sigt.

Der er idag et bredt funderet Life Science forskningsmiljø i Danmark som er langt større end Danmarks relative størrelse svarer til både på grund af den store danske farmaceutiske industri og tidlige satsninger

på bioinformatik på flere danske universiteter. En af de centrale udfordringer er dataeksplosionen med data fordoblingstid på 5 måneder, der gør at der skal udvikles nye metoder og modeller da nye data typisk sammenlignes med en stor del af de eksisterende data. Hvis Danmark skal fastholde sin succes i den internationale konkurrence, skal vi "poole" ressourcer. Et nationalt Life science HPC center vil skabe et framragende udgangspunkt herfor.

Den danske indsats koordineres af Center for Biologisk Sekvensanalyse på DTU (CBS), der har opbygget en betydelig ekspertise og knowhow indenfor anvendelse og tilrettelæggelse af HPC for Life science over de sidste 20 år. CBS og de andre danske ELIXIR partnere har sammen opbygget ekspertise indenfor heterogene softwaremiljøer og datatyper som karakteriserer Life Science forskeres brug af HPC. CBS er internationalt anerkendt for sine tools samt prediction servere (som er udviklet af forskere der idag også er ansat på flere andre danske universiteter), og har flere millioner hits på websitet www.cbs.dtu.dk om måneden, hvilket har stor betydning for de danske universiteters ranking og synlighed.

I det danske ELIXIR samarbejde arbejder vi med at:

- Birdrage til en transnational infrastruktur for biologisk information og relateret serviceorienteret virksomhed, der dækker eksisterende nationale infrastrukturer og netværk.
- Fremme brugen af IT-teknologi i forbindelse med dataintegration og databasekompabilitet.
- Fremme og videreudvikle brugen af distribuerede annoteringsteknologier for store biologiske databaser.
- Øge kompetencen og størrelsen af den allerede store skare af brugere ved at styrke nationale tiltag omkring uddannelse.
- Øge effektiviteten af dansk samarbejde gennem forbedret dataudveksling.
- Etablere links mellem molekylære databaser og udvikle ressourcer for medicinsk forskning (f.eks. biobanker), landbrug og miljø (f.eks. biodiversitet).
- Skabe infrastruktur der muliggør koordineret brug af værktøjer til dataanalyse (f.eks. v.h.a. web-service teknologi).

Innovation og kommercialisering. Ud over de kendte typer af anvendelser, hvor bioinformatiske computeranalyse danner grundlag for udvikling af biologisk/biokemisk indsigt (og på den måde medfører at de danske, industrielle biologiskmedicinske forskningsmiljøer får endnu mere ud af de resourser de allerede har), er der et stort forretningsmæssigt potentiale i at virksomheder systematisk bliver bedre til at anvende den nyeste IT-viden. Indenfor den bioteknologiske industri er der en lang række IT-områder, hvor der kan forudses en udvikling drevet af bioinformatisk forskning, snarere end af eksperimentel forskning. Det vil typisk være tilfældet, når den gængse teknologi inden for et felt ikke indfanger de problemstillinger, der findes i forbindelse med anvendelsen af teknologien i bioinformatisk sammenhæng. Dette gælder f.eks. inden for netværks-baseret beregning af vekselvirkningen mellem gener, proteiner og metabolitter i en celle. Her kan man forvente at en integration af optimeringsmetoder og simulering vil give nye muligheder.

Videreudvikling af metoder til effektiv dataorganisering og udnyttelse af eksisterende avancerede datastrukturer kommer til at spille en central rolle i forbindelse med data-mining.

Både selve bioinformatikforskningen, men særligt alle de områder der er dens aftagere indenfor biologien, bioteknologien og den medicinske forskning, har et meget højt kvalitetsniveau i Danmark (f.eks. målt på citationer), og bidrager afgørende til landets samlede forskningsimpact. Den danske bioinformatikinfrastruktur har links til mange andre forskningsinfrastrukturer, herunder biobankerne, registre, data-grids, imaging infrastrukturer o.s.v.

Peter Løngreen, Senior Scientific Consultant, ELIXIR Denmark coordinator

Center for Biological Sequence Analysis

Department of Systems Biology, Technical University of Denmark

Building 208, 2800 Kgs. Lyngby

Søren Brunak, prof., PhD, ELIXIR grant responsible

Center for Biological Sequence Analysis, Center director

Dept. of Systems Biology, Technical University of Denmark,

Building 208, 2800 Lyngby,

and

Novo Nordisk Foundation Center for Protein Research

Programme Director

Disease Systems Biology, Faculty of Health Sciences, University of Copenhagen,

Blegdamsvej 3A, 2200 Copenhagen, Denmark