Århus, den 26. marts 2008

Simple matematiske modeller til beskrivelse af komplekse biologiske systemer

Carsten Wiuf

Center for Bioinformatik (BiRC) ved Aarhus Universitet

Biologien er i disse år i rivende udvikling blandt andet på grund af de mange tekniske fremskridt der gør det muligt simultant at måle og registrere tusinder af variable. For eksempel er det i dag muligt at bestemme aktiviteten af alle gener i en celle på en gang, hvilke af en celles tusinder af proteiner der vekselvirker med hinanden eller afkode millioner af variable DNA-positioner i det humane genom ved et enkelt eksperiment.

Den slags målinger giver os mulighed for at studere celler, væv og organismer i et nyt perspektiv – et helheds- eller systemperspektiv, hvor tendenser og mønstre træder frem som ikke er synlige, hvis kun en lille del af systemet observeres. Således kan man for eksempel se, at generne i en hjernecelle er markant anderledes udtrykt end generne i en levercelle, og at udtrykkene ændrer sig gennem en organismes levetid. Udover at være interessant i sig selv, åbner sådanne observationer for nye muligheder inden for fx lægevidenskaben, idet vi ud fra et systemperspektiv kan sammenligne syge celler med raske og belyse sygdommes årsager og undersøge effekten af medicin i et nyt lys.

Ikke kun den molekylære biologi er i rivende udvikling, men også andre grene af biologien oplever tilsvarende forandringer i form af øgede datamængder, øget information og viden. Dette stiller naturligvis krav til måden at registrere og indsamle data på, samt til hvordan vi efterfølgende behandler og analyserer data med henblik på spørgsmål vi finder videnskabeligt eller samfundsmæssigt relevante. Derfor kan man parallelt med udviklingen inden for biologien spore en

udvikling inden for statistik, datalogi og matematik, hvor man søger at udvikle modeller, metoder og teknikker til håndtering og analyse af store datamængder; dette kræver indsigt i biologisk teori. Det samspil eller krydsfelt, der dermed opstår mellem biologi, statistik, datalogi og matematik kaldes for bioinformatik.

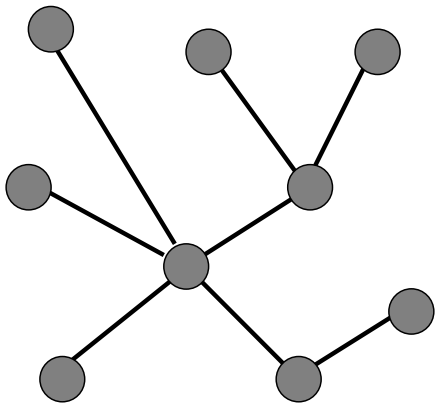
Faget bioinformatik har udviklet sig fra i slutningen af 60erne at være enkeltpersoners sammenligninger af små DNA- og proteinsekvenser til i dag at være centre med mange ansatte med forskellige baggrunde og ekspertiser. I Danmark er der større centre (>20 ansatte) på Danmarks Tekniske Universitet, Københavns Universitet og Aarhus Universitet. Jeg er selv ansat på Center for Bioinformatik ved Aarhus Universitet og her er folk med baggrund i matematik, statistik, datalogi, biologi, kemi, molekylærbiologi og ingeniørvidenskab. Dette giver et meget spændende arbejdsmiljø, hvor den fælles interesse er at forstå biologiske problemstillinger gennem teoretiske overvejelser.

Selvom alle de nævnte fagområder gennem årtier har opereret med begrebet *systemteori* i forskellige former, er begrebet dukket op på ny i forøget styrke som en følge af de muligheder, der har vist sig inden for biologien i dag. Her vil jeg beskrive et biologisk system kaldet *interaktomet* og vise hvordan vi med simple matematiske modeller kan beskrive hvordan interaktomet gennem evolution har formet sig over tid. Artiklen er bygget op så jeg først beskriver biologien, dernæst motiverer valget af matematisk model og diskuterer en statistisk analyse af biologiske data vha. modellen. Endelig fortolker jeg de statistiske resultater til matematiske udsagn om modellen.

Interaktomet: Biologisk fortolkning og evolution

Mængden af alle de forskellige proteiner i en organisme samt de fysiske bindinger, der er mellem par af proteiner kaldes for interaktomet; se Figur 1. Proteiner er produkter af gener, og et gen

udøver til dels sin funktion gennem dets protein og de vekselvirkninger proteinet kan have med andre proteiner. Det enkelte protein påvirker cellens processer ved at binde til andre proteiner, gener og molekyler og interaktomet repræsenterer dermed den del af cellens processer, der er bestemt af bindinger mellem proteiner.



Figur 1. Interaktomet illustreret som en graf med proteiner som knuder og fysiske vekselvirkninger som kanter. I eksemplet er der N=9 knuder og M=8 kanter. Graden af en knude er antallet af kanter der forbinder den med andre knuder.

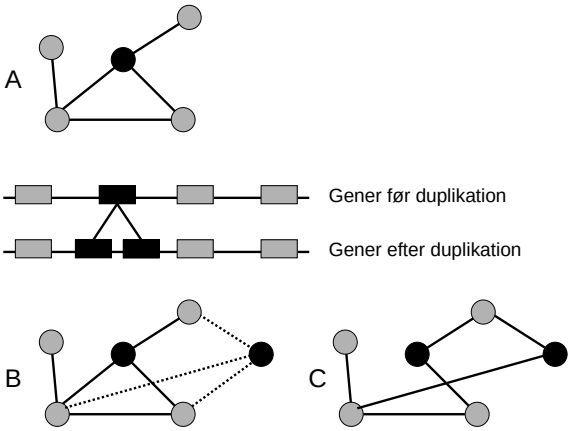
Der er derfor stor interesse for at forstå interaktomet og de evolutionære processer der har skabt det. Et eksperimentelt bestemt interaktom kaldes et *proteininteraktionsnetværk* (PIN). I dag er PIN-data fra en række organismer tilgængelige, dog er PIN-data fra mennesket endnu sparsomme, da de anvendte eksperimentelle teknikker stadig er dyre og tidskrævende. Desværre er PIN-data også ufuldstændige i den forstand at kun kendte proteiner kan medtages i et eksperiment. I Tabel 1 er størrelsen på typiske PIN-datasæt angivet med et bud på hvor fuldstændigt datasættet er.

Organisme	Knuder	Kanter	Total	Procent
S. cerevisiae (gær)	4.959	17.226	5.500	90
D. melanogaster (flue)	7.451	22.636	12.900	58
C. elegans (orm)	2.638	3.970	22.000	12
H. pylori (bakterie)	675	1.096	1.500	45
P. falciparum (parasit)	1.271	2.642	5.300	24

Tabel 1. Antallet af knuder og kanter i typiske PIN-datasæt. Total: Estimeret interaktomstørrelse. Procent: Antallet af knuder (proteiner) i % af det estimerede antal af proteiner i organismen. Data er taget fra en offentlig tilgængelig database DIP (Database of Interacting Proteins).

En organismes interaktom udvikler sig evolutionært over tid. I sin yderste konsekvens forestiller vi os at interaktomet er startet som et enkelt protein og vokset i størrelse over tid. Da vi ikke til fulde forstår livets udvikling i dets tidlige faser, er interaktomets udvikling fra et enkelt protein at betragte som en bekvem abstraktion. Til gengæld findes en række andre biologiske processer, der ændrer interaktomet, og som vi forstår langt bedre. Den vigtigste er genduplikation, en proces hvorved et eksisterende gen bliver kopieret og indsat i genomet, så der lige efter duplikationen er to kopier af det samme gen. Disse to gener er identiske og producerer derfor det samme protein. Interaktomet ændrer sig altså ikke umiddelbart som følge af en genduplikation. Over relativt kort tid (i evolutionær forstand) udvikler de to gener sig; her er der forskellige muligheder: 1) Det ene gen ødelægges af skadelige mutationer og ophører med at være aktivt, 2) Det ene gen får en helt ny funktion vha. gavnlige mutationer eller 3) De to gener deler det oprindelige gens funktioner mellem sig. Den sidste mulighed menes at være langt den vigtigste. De ændrede funktioner giver anledning

til ændringer i de proteiner generne producerer og dermed også til ændringer i de bindinger proteinerne har til andre proteiner. Dvs. over tid sætter genduplikation sig spor i interaktomet; se Figur 2.



Figur 2. Figuren viser genomet før og efter en genduplikation og det tilhørende proteinnetværk (A). Lige efter duplikationen af det sorte gen har de to sorte geners proteiner de samme interaktioner (B), men i løbet af evolutionær kort tid differentieres disse (C). I den matematiske model antages at begge kanter til et gråt protein bevares med sandsynlighed p og at en kant skabes mellem de to sorte proteiner med sandsynlighed q. Hvis kun en kant bevares, vælges med lige sandsynlighed. I eksemplet er der sandsynlighed $p(1-p)^2(1-q)/4$ for at observere netværket i (C), hvis det sorte gen duplikeres.

Genduplikation er evolutionært en meget vigtig proces, idet den giver et gen/protein mulighed for at specialisere sig (se punkt 3) – fx ser man ofte proteiner der kun er udtrykt i en bestemt type væv, men som er tæt beslægtede med andre proteiner med lignende funktioner i andet væv. Derudover findes der en række andre processer, der specielt er vigtige i bakterieevolution og som vi samlet vil betegne *tilknytningsprocesser.* De dækker integration af gener fra vira og andre bakterier og fusion af eksisterende gener i genomet, hvorved organismen bliver i stand til at udvikle gener, og dermed proteiner med helt nye funktioner. Den relative hyppighed af og vigtighed mellem de forskellige typer processer er i det store hele ukendt.

En matematisk model af interaktomets evolution

Første del af den matematiske model består i at repræsentere interaktomet ved hjælp af en graf som vist i Figur 1. Grafer (eller netværk) finder i dag hyppig anvendelse i mange naturvidenskabelige fag blandt andet fordi der er udviklet en rig og sund matematisk teori for grafer. De er desuden nemme at visualisere og derfor naturligt appellerende.

De grafer vi betragter ændrer sig over evolutionær tid på en måde vi ikke på forhånd kender, og vi beskriver dem derfor i sandsynlighedsteoretiske eller stokastiske termer. Lad $G_t = (V_t, E_t)$ betegne en graf med t knuder, hvor V_t er mængder af knuder og E_t er mængden af kanter i grafen. Et netværk bygges op således:

- Start med en graf $G_s = (V_s, E_s)$ af størrelse s. Vi vælger s=1, dvs. en knude uden kanter
- Antag grafen i skridt t er $G_t = (V_t, E_t)$. I næste skridt, t+1, gør følgende:
    - Duplikation: Vælg med sandsynlighed $0 \leq \alpha \leq 1$ en knude tilfældigt i grafen og kopier knuden med alle dens kanter; se Figur 2. Dernæst modificeres den nye og den gamle knude vha. parametrene p og q, som beskrevet i figur 2
    - Tilknytning. Vælg med sandsynlighed $1-\alpha$ en knude tilfældigt i grafen. Forbind en ny knude til den valgte knude ved hjælp af en enkelt kant
- Fortsæt indtil netværket har den ønskede størrelse. Dernæst udvælges knuder så antallet passer med et observeret PIN-datasæt

Der er naturligvis mulighed for variationer og for at udvikle modellen yderligere. Den beskrevne model er specificeret ved tre parametre (α, p, q), og da det er en evolutionær model vil grafen i skridt t+1 kun afhænge af grafen i det foregående skridt t – dvs. vores grafgenerende proces er en

Markov-kæde. Desuden er enhver modifikation af grafen lokal – dvs. den involverer kun få knuder (den nye og den gamle, plus den gamles naboer). Disse egenskaber sætter os i stand til at udsige interessante udsagn om de grafer vi kan generere.

To eksempler: *Helicobacter pylori* og *Plasmodium falciparum*

Jeg og min gruppe har analyseret PIN-datasæt fra *H. pylori* (675 knuder) og *P. falciparum* (1.271 knuder). I øjeblikket er vi ved at analysere et PIN-datasæt fra gær (*Saccharomyces cerevisiae*) med ca. 4.500 knuder, hvilket er på grænsen af hvad vi i øjeblikket beregningsmæssigt kan klare inden for en overkommelig tid.

H. pylori er en lille bakterie der kan forårsage mavesår og derfor har været genstand for en del forskning. Der findes medicin mod *H. pylori*, men flere resistente bakteriestammer er kendte og den er desuden mistænkt for at spille en rolle i udviklingen af cancer. *P. falciparum* er en encellet parasit, der forårsager malaria i mennesker. Siden man i 1976 fik held til at dyrke *P. falciparum* i laboratorier er malariamedicinen blevet betragteligt forbedret, men sygdommen kræver stadig flere millioner menneskeliv om året på verdensplan. Resultaterne fra vores analyser er ganske interessante og opsummerede i Tabel 2.

	p	q	α	Kanter
H. pylori	0,56	0,05	0,78	5.636
P. falciparum	0,51	0,05	0,93	43.835

Tabel 2. Vist i tabellen er estimater for p, q og α. Kanter er det estimerede antal kanter i det fulde (uobserverede) interaktom. Antallet af knuder og kanter i de to PIN-datasæt fremgår af Tabel 1. Estimaterne er opnået vha. avancerede statistiske metoder (primært Bayesian Statistics).

For det første ser vi at de to parametre der bestemmer genduplikation er meget ens i de to datasæt. Sandsynligheden (q) for at den nye og den kopierede (gamle) knude er forbundet er meget lille, hvilket er konsistent med hvad andre har fundet ved analyse af andre typer datasæt. Det peger også på at genduplikationens rolle er at skabe nye funktioner og ikke blot at modificere de gamle.

Derimod er sandsynligheden (p) for at både den nye og den gamle knude bevarer en kant til en knude ca. fifty-fifty og den ligger tæt på den grænse, hvor netværket under nogen omstændigheder ophører med at være stabilt (se næste afsnit). Endelig ser vi at parameteren α, der angiver hvor hyppigt genduplikation sker i forhold til tilknytninger, er meget forskellig i de to datasæt. *H. pylori* er som nævnt en bakterie, og bakterier udveksler ofte genetisk materiale mellem hinanden ved hjælp af de processer, der her er modelleret som tilknytninger. Derimod er *P. falciparum* en encellet eukaryot (den er grundlæggende opbygget som vores celler) og i eukaryoter er den slags processer langt sjældnere. Dvs. resultaterne stemmer overens med vores forventning.

Endvidere er de to organismers interaktomer også meget forskellige i størrelse, hvilket naturligt hænger sammen med hvor mange proteiner de hver især har (Tabel 1). Antallet af mulige kanter vokser kvadratisk med antallet af knuder, $N(N\text{-}1)/2$, og ratioen $(2M)/[N(N\text{-}1)]$ af faktiske kanter M til antallet af mulige kanter kan opfattes som udtryk for netværkets forbundenhed. I tilfældet *H. pylori* er ratioen ca. $5 \cdot 10^{-3}$ mens den er ca. $3 \cdot 10^{-3}$ for *P. falciparum*, dvs. bakterien er mere forbundet end eukaryoten. Også dette er i tråd med vores forventning, idet eukaryoten har mange varierede funktioner, der varetages af forskellige proteiner, mens bakteriens proteiner hyppigere er involveret i flere forskellige funktioner.

Matematiske resultater

Vi kan bevise en række udsagn om denne simple model. Endvidere bemærker vi at angivelsen af modellen svarer nøje til hvordan man kan simulere tilfældige grafer fra modellen, og at vi kan lære om modellen ved at simulere grafer og registrere deres egenskaber. Her vil vi kun interessere os for graden af en knude, dvs. hvor mange andre knuder den er knyttet til. Man kan opstille følgende rekursion for det forventede antal knuder $n_t(k)$ af grad k i grafen G_t:

$$n_{t+1}(k) = \left(1 - \frac{1+kp}{t}\right) n_t(k) + \frac{1+(k-1)p}{t} n_t(k-1) + 2\sum_{j \geq k-1} \binom{j}{k-1} \psi^k (1-\psi)^{j-k+1} \frac{n_t(j)}{t}$$

hvor $\psi=(1+p)/2$ og vi for simpelheds skylds har antaget at $q=1$ og $\alpha=1$ (helt tilsvarende resultat kan opnås generelt blot med en del flere led).

Det første led kommer fra at betragte en knude A af grad k og beregne sandsynligheden for at den bevarer graden efter en duplikation; det sker, hvis ingen af dens naboer eller den selv udvælges (sandsynlighed $1-[1+k]/t$) eller hvis en af naboerne udvælges til duplikation, men ikke etablerer en kant til A (sandsynlighed $[1-p]k/t$). Tilsvarende fremkommer andet led ved at betragte en knude af grad $k-1$ og beregne sandsynligheden for at dens grad stiger med 1. Endelig angiver sidste led graden af den nye knude og knuden der duplikeres ('den gamle'): hvis den gamle knude har grad j får den nye og den gamle knude hver et binomialt fordelt antal kanter fra naboerne, samt en kant i mellem dem ($q=1$). Rekursionen er en konsekvens af Markov-egenskaben omtalt tidligere.

Rekursionen kan kun løses eksplicit i trivielle tilfælde, men det er i nogle tilfælde muligt at afgøre om hyppigheden af knuder af grad k stabiliseres som netværket vokser, eller om der ikke indtræder en ligevægt i takt med at netværket bliver vilkårligt stort. For et vilkårligt parametervalg (α, p, q) har vi:

- Hvis $\alpha p<0{,}5$ findes en ligevægt (ergodisk rekurrent).
- Hvis $\alpha=1$ og $p<0{,}533$ har et uendelig stort netværk *uendelig* mange knuder af vilkårlig grad, men en ligevægt findes ikke nødvendigvis (rekurrent). Her er 0,533 den approksimative løsning til ligningen $2\log(\psi)+p=0$.
- Hvis $\alpha=1$ og $p>0{,}562$ har et uendelig stort netværk højst *endelig* mange knuder af en vilkårlig grad, dog potentielt uendelig mange af grad nul (transient). Her er 0,562 den approksimative løsning til ligningen $1/(1-\psi)+1/(p+\psi)-p-2=0$.
- Hvis $\alpha<1$ har et uendelig stort netværk *uendelig* mange knuder af vilkårlig grad, men en ligevægt findes ikke nødvendigvis (rekurrent).

Hvad der sker i det lille vindue mellem 0,533 og 0,562 ved vi ikke, ej heller om grænsen for en ligevægt er højere end 0,5. Simulationer er af lille hjælp her for selv hvis vi simulerede meget store netværk, ville vi ikke være i stand til at afgøre om graderne stabiliseres eller ej. Det første resultat hænger tæt sammen med den gennemsnitlige grad af en knude der er

$$\frac{2-2(1-q)\alpha}{1-2\alpha p}$$

forudsat $\alpha p<0{,}5$ og ellers uendelig.

Biologisk er resultaterne interessante, idet de fortæller os om hvad vi skal forvente for et virkeligt netværk. Bringer evolutionen graden af en knude til en ligevægt? Ved hjælp af matematiske overvejelser bliver det således muligt at klassificere systemerne efter deres stabilitet. I de to tidligere eksempler finder vi således at de begge falder under betingelsen $\alpha<1$, men også at parametrene for *H. pylori* ligger nær stabilitetsgrænsen (både α og p).

Afsluttende bemærkninger

Da modellen på flere måder er en tilnærmelse til virkeligheden, skal vi være påpasselige med at fremhæve resultaterne ukommenterede og man bør undersøge hvorvidt resultaterne er robuste over for forskellige modifikationer af modellen. Det er dog meget interessant at estimaterne stemmer fint overens med vores forventning og viden fra andre typer data. Desværre er stadig kun få PIN-datasæt tilgængelige – med tiden kommer der forhåbentlig flere til. Nogle af de metodiske spørgsmål der trænger sig på er, hvorledes man kan analysere PIN-data fra tæt beslægtede arter, således at man tager højde for deres fælles evolution, og hvordan man kan analysere data fra arter der har samspillet gennem evolution. Et eksempel på sidstnævnte er parasitten *P. falciparums* samspil med mennesket. I den forbindelse vil det være interessant at betragte farvede netværk, således at knuderne fx kan antage farve efter deres funktion. Det er sandsynligvis kun proteiner med visse funktioner i *P. falciparum*, der har tilpasset sig mere menneskelige omgivelser.

I denne artikel har jeg forsøgt at give et indblik i nogle af de problemstillinger, data og metoder moderne biologer og ”anvendte” matematikere tumler med. Det er blevet spået at den anvendte matematik inden for biologi vil gennemgå en eksplosion næsten af samme størrelse som biologien selv har gennemgået. Det bliver næppe tilfældet, men interessant er det i hvert fald at opleve hvordan mange biologiske problemstillinger kan hjælpes godt på vej af matematik og at mange unge matematikere er begyndt at søge mod den anvendte matematik. Som ”anvendt” matematiker hilser jeg naturligvis dette velkomment.

Efterskrift

Flere af mine studerende og postdocs har arbejdet med netværk og en speciel tak går til Oskar Hagberg, Oliver Ratmann og Michael Knudsen, som alle har ydet væsentlige bidrag. Jeg ønsker

også at takke Michael Knudsen og Freddy Bugge Christiansen for kommentarer til manuskriptet og Enette Berndt Knudsen for hjælp med det stilistiske.

Litteratur i udvalg

Desværre er jeg ikke bekendt med populærvidenskabelige fremstillinger af emnet, dog er 'Linked' af Barabasi en glimrende populær introduktion til den omsiggribende netværkstænkning. Den videnskabelige litteratur er rig og omfatter bidrag af matematisk, statistisk, fysisk og biologisk karakter. Nedenfor er udvalg med vægt på nogle af mine egne bidrag.

A.-L. Barabasi (2003) Linked, Penguin Books.

R. Durrett (2006) Random Graph Dynamics, Cambridge University Press.

M. Lynch (2007) The Origins of Genome Architecture, Sinauer Press.

S. Ohno (1970) Evolution by Gene Duplication, Springer Verlag.

M. Knudsen og C. Wiuf (2008) A Markov chain approach to Randomly Grown Graphs, Journal of Applied Mathematics 2008: 190836.

O. Ratmann *et al.* (2007) Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*. PLoS Computational Biology 3: e320.

C. Wiuf *et al.* (2006) A likelihood approach to analysis of network data. Proceedings of the National Academy of Science USA 103: 7566-7570.

DIP, http://dip.doe-mbi.ucla.edu/ (proteininteraktionsdata).