Bioinformatik – en statistisk disciplin

Af: Niels Richard Hansen
Afdeling for Anvendt Matematik og Statistik, KU
Email: richard@math.ku.dk

Fra beregningsbiologi til bioinformatisk statistik

Op gennem halvfjerdserne og firserne gav udviklingen indenfor biologien anledning til nye, spændende og tværfaglige problemer. Særligt indenfor forskningen i det molekylære grundlag for livet opstod der et behov for at organisere, håndtere og anvende mere og mere omfattede molekylærbiologiske data og ikke mindst at lave beregninger med disse data.

En af de oprindelige problemstillinger drejer sig om at sammenligne proteiner, der, som det funktionelle produkt af vores gener, udgør en yderst vigtig klasse af makromolekyler. Proteiner er store molekyler bygget op som lineære sekvenser af aminosyrer, og en metode til sammenligning går ud på at opfatte proteinmolekylerne som simple tekststrenge fra et 20 bogstavs alfabet (de 20 forskellige aminosyrer). Dernæst placeres de to tekststrenge over hinanden, og man prøver, med tilladelse til at indsætte mellemrum, at "skubbe" bogstaverne i de to sekvenser frem og tilbage, indtil aminosyrerne er parrede, så de "passer" godt sammen. Denne metode kaldes ofte på engelsk for alignment – et ord, der vist ikke har en god dansk oversættelse.

```
ACAD-ACDADCCA
ADADCACAADDCA
```

Figur 1: Et eksempel på en alignment af to proteiner. Der er indsat et mellemrum i den øverste sekvens af aminosyrer, og der er tre ikke-matchende aminosyrer. I eksemplet benyttes kun de tre aminosyrer `A` (Alanin), `C` (Cystein) og `D` (Asparaginsyre).

En sådan sammenligning af proteiner giver bl.a. et værktøj til at studere evolution på det mest fundamentale niveau – det molekylære. Man forestiller sig, at to evolutionært relaterede proteiner også ligner hinanden aminosyre for aminosyre, og at en sammenligning derfor kan vise, om to proteiner faktisk er relaterede. En del af anstrengelserne gik i starten ud på at løse rent beregningsmæssige problemer og var således meget algoritme-orienteret. Området gik under navnet *computational biology* – eller direkte oversat *beregningsbiology*. Siden hen har betegnelse *bioinformatik* slået an som en slags paraplybetegnelse for arbejdet med og bearbejdning af især molekylærbiologiske data ved hjælp af en computer. Bioinformatik er som område blevet et sammenrend uden lige – absolut forstået på den gode måde – af forskere med vidt forskellige baggrunde såsom biologi, datalogi og statistik samt fysik og kemi, og forskningen spænder bredt fra de egentlig biologiske spørgsmål, der kræver bioinformatiske værktøjer, over en masse databasearbejde til deciderede teoretiske metodespørgsmål. Det er primært i den sidste kategori, at den statistiske forskning bidrager til udviklingen af bioinformatikken.

Evolutionsmodeller for proteiner

For det klassiske problem, sammenligning af proteiner, starter den statistiske modellering med evolutionshypotesen på molekylært niveau. Den går ud på, at aminosyrer enkeltvis eller i grupper muterer, således at nye funktioner kan opstå. Langt de fleste mutationer er katastrofale, men enkelte vil overleve og blive leveret videre til kommende generationer. Proteiner, der har en ikke for fjern fælles fortid, vil ligne hinanden, selv om de er forskellige. Men hvordan vurderer man, om en konkret (god?) alignment er udtryk for andet end tilfældige ligheder mellem proteinerne? Det statistiske modelkoncept giver her en fornuftig angrebsvinkel. Det man skal gøre er at opstille en model dels for mutationsprocessen og dels for ikke-relaterede proteiner. For to givne proteiner bør enhver beslutning om, hvorvidt de er relaterede, baseres på en sammenligning af sandsynligheden for at observere proteinerne som følge af en mutationsproces fra et fælles forfaderprotein, med sandsynligheden for at proteinerne er ikke-relaterede. Man kan godt opfatte dette som et statistisk test af relaterethed mod ikke-relaterethed, men det er bedre at tænke på problemstillingen som et klassifikationsproblem. Den statistiske modellering giver dernæst en mulighed for at diskutere, hvad fejlfrekvensen er i det lange løb, dvs. når vi klassificerer mange par af proteiner som relaterede/ikke-relaterede, hvor mange fejl begår vi så i gennemsnit. Men bemærk, at de modeller vi har valgt, meget vel kan være forsimplede og evt. forfejlede i forhold til den virkelige evolutionsproces. Den statistiske teori viser, at proceduren beskrevet ovenfor essentielt er den korrekte (optimal i en rimelig forstand), vel at mærke *hvis modellerne er korrekte*. Skal vi forbedre vores klassifikationsprocedure, skal vi altså finde nogle bedre modeller. Det er i den henseende tankevækkende, at de modeller, som kan behandles og benyttes i praksis, kun kan håndtere de mest rudimentære evolutionære hændelser, såsom substitution af enkelte aminosyrer. Det burde dog blot give blod på tanden – der er masser at komme efter i denne branche!

Men hvordan modellerer man så mutationsprocesser? Den klassiske model er givet ved, at hver enkelt aminosyre muterer i henhold til en tidshomogen markovproces, og at aminosyrerne muterer uafhængigt af hinanden. Den store fordel ved denne model er, at man kan beskrive mutationsprocessen med relativt få parametre, nemlig 380 mutationsintensiteter. Grundet tidshomogeniteten giver dette en model for relaterede proteiner – uafhængigt af hvor langt tilbage i fortiden vi skal for at finde en fælles forfader. Modellen giver endvidere anledning til PAMn-scoring for $n \geq 1$, jf. tabel 1, som ofte bruges til scoring af parrede aminosyrer, når man aligner proteiner. Her er n den evolutionære afstand mellem proteinerne målt i såkaldte PAM-enheder. Et nyere alternativ er BLOSUMn-scoring for $1 \leq n \leq 100$, som ligeledes er baseret på en model for mutation, hvor de de enkelte aminosyrer muterer uafhængigt af hinanden, men hvor der ikke er nogen

antagelse om en underliggende tidshomogen markovproces. Her betegner n frekvensen af det totale antal konserverede, altså ikke-muterede, aminosyrer i proteinet. F.eks. for $n = 45$ matcher 45% af aminosyrerne fra de to proteiner. Det er den gængse opfattelse, at BLOSUM-scoring fungerer lidt bedre end PAM-scoring.

	PAM250			BLOSUM45		
	A	C	D	A	C	D
A	2			5		
C	-2	12		-1	12	
D	0	-5	4	-2	-3	7

Tabel 1: Udsnit af to hyppigt anvendte metoder til scoring af parrede aminosyrer i en alignment af to proteiner. PAM250 og BLOSUM45 er fortolkningsmæssigt sammenlignelige, men baseret på to forskellige modeller.

Bruger man PAM250-scoring på den alignment vi ser i figur 1 (læg bidragene for alle de parrede aminosyrer sammen) finder vi en samlet score på 32. Den tilsvarende BLOSUM45-score for den given alignment er 55.

Datamining og Molekylære strukturer

En af de begivenheder, som for alvor satte fut i bioinformatikken, var beslutningen om, at kortlægge det menneskelige genom. Det er i dag lykkes, og kortlægningen af genomet for en række andre organismer er også afsluttet eller i fuld gang. Det har betydet fri adgang til enorme mængder rå genom-data, som bioinformatikere i hele verden gennemtrawler dag efter dag for at finde ny og dybere forståelse for de grundlæggende livsprocesser. Man vil f.eks. gerne forstå, hvordan hele symfoniorkestret af gener spiller sammen. Biokemien giver os god indsigt i, hvordan de enkelte instrumenter – vel og mærke dem vi har opdaget – spiller alene, men vi forstår stadig meget dårligt, hvordan det hele spiller sammen. Derfor prøver man med en bred vifte af *datamining* værktøjer at undersøge genomer på kryds og tværs for at finde *alle* instrumenterne, så man bedre kan få et komplet billede af orkestret. Fælles for en meget stor del af sådanne datamining værktøjer er, at de er baseret på statistiske modeller.

Min egen forskning har bl.a. beskæftiget sig med modeller og datamining teknikker, som tager udgangspunkt i 3-dimensionale molekylære strukturer. Dvs. teknikker til at søge i genomer efter sekvenser, som i 3 dimensioner vil folde sig sammen til specifikke strukturer. Det er som hovedregel strukturen snarere end den lineære tekststreng, som bestemmer molekylernes funktion, og derfor er det yderst relevant at modellere netop strukturen. Det er imidlertid noget vanskeligere end at modellere tekststrenge, og man løber hurtigt ind i problemer såsom voldsomt mange ukendte parametre og exceptionelt langsommelige algoritmer. Mit arbejde har centreret sig om at modellere en relativt simpel type struktur kaldet en stem-loop, og om det teoretiske grundlag for at beregne fejlfrekvenser, når man søger efter stem-loops i genomer. Der er en vældig spændende og udfordrende teori at udforske i forbindelse med disse beregninger, og i matematisk sofistikerethed lader teorien ikke noget tilbage at ønske.

Microarrays

Man må dog erkende, at uanset den store triumf kortlægningen af det menneskelige genom er, så giver kortet (i sin nuværende form) kun et begrænset og passivt billede af genomet. Det, man har kortlagt, er den rå og til dels ufordøjede information, man finder placeret i kromosomerne, og som vi arver fra vores forældre. Det er således arkitekttegningen, vi har fat i, men en samlet forståelse af, hvordan det hele udvikler sig dynamisk i hver enkelt celle, er vi stadig lysår fra. Datamining teknikker eller ej, så skal der noget mere til, hvis vi rigtigt skal forstå biologien. Her kommer en anden og ret ny eksperimentel teknik ind i billedet. En teknik, som især har fået statistikere op af stolen og til at kaste sig over bioinformatikken. Ved hjælp af såkaldte *microarrays* får vi et slags dualt, dynamisk billede af genomet. Microarrays er små plader inddelt i endnu mindre celler, hvor man i hver celle har mulighed for at aflæse hvor stor en mængde protein fra et givet gen, der faktisk er udtrykt, altså produceret, i cellen til et givet tidspunkt. Man kan f.eks. benytte denne teknik, til at se hvad der sker på gen-niveau, når to ellers identiske celler udsættes for forskellige behandlinger. Herved kan man lære, hvordan de mange gener coreguleres under forskellige omstændigheder. Men der er mange, mange andre anvendelsesmuligheder af microarrays, og kun fantasien synes at sætte grænser.

Når det nu er sagt, så er der også en del problemer. Microarrays er teknisk set ikke så lige til, og der er et antal lag, hvor der kan opstå fejl og støj af forskellig slags, udover de rent biologiske fluktuationer man må forvente. F.eks. måles mængden af protein ikke direkte, men via mellemproduktet messenger-RNA. Principielt adskiller modelleringen af disse fejl sig ikke så meget fra mange andre klassiske statistiske modeller, hvor man observerer en flerdimensional, kvantitativ størrelse, og hvor man ønsker at modellere sådan noget som niveau, variation og co-variation. Men det er dog en væsentlig udfordring, at hver observation fra et microarray giver en datavektor, som ofte er 10.000-dimensional eller mere, og at man sjældent har særligt mange arrays at gøre godt med (det er en dyr teknik, og der er ikke altid konsensus om, hvordan pengene skal bruges). Endvidere er det nødvendigt, at man sætter sig ind i en del af den molekylære biologi/kemi og andre tekniske detaljer om, hvordan microarrays faktisk fungerer – og hvad det er for nogle biologisk videnskabelige spørgsmål, der er interessante – for at man kan konstruere rimelige modeller.

Og alt det andet...

Der er selvfølgelig mange andre bioinformatiske emner af statistisk karakter, som jeg af gode grunde ikke kan komme ind på. Ingen nævnt, ingen glemt.

Som statistiker ser jeg bioinformatikken som et rigtigt godt område at kaste sig over. Der er for det første fokus på området, som er i vældig vækst. Der er mange spændende teoretiske metode- og modelspørgsmål, som venter på at blive taget op. Og så er der adgang til omfattende databaser med bunker af data, der kun venter på at blive analyseret. For slet ikke at tale om den sidegevinst, at man får en indsigt i den meget spændende, biologiske forskning, der foregår i øjeblikket.