RNA er jo bare matematik!

Hvordan kan man kurere sygdomme med matematiske geometriske strukturer? Det kan man i princippet, hvis de geometriske figurer er RNA-molekyler, og sygdommen skyldes syge gener

Af Jakob Blaavand

■ I bund og grund handler det om at forudse molekylet RNAs geometriske udformning, på baggrund af kendskabet til rækkefølgen af de byggeklodser (nukleotider), det er opbygget af. RNA-molekylet kan i princippet bruges til at kurere sygdomme med en teknik, der hedder antisense. *Antisense*-teknikken er essentielt en snedig måde at slukke for gener, der er årsag til sygdomme. Ved at slukke for sådanne syge gener, kan man forhindre sygdommen i at udvikle sig.

Ved Centre for Quantum Geometry of Moduli Spaces (QGM) ved Aarhus Universitet har forskere sammen med kolleger fra Kina og Tyskland, udviklet metoder, der kan forudsige godt 90 % af geometrien af korte RNA-sekvenser.

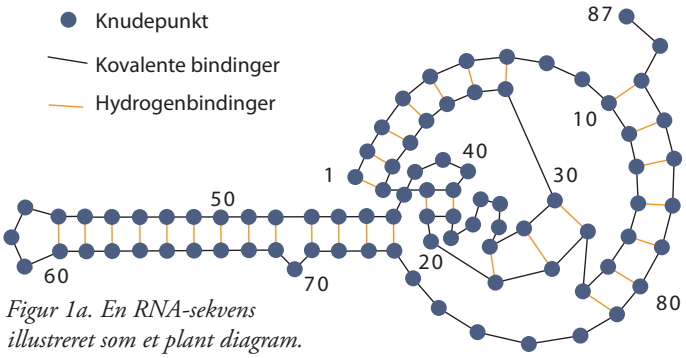
Forudsigelser af RNA'ets form

RNA-molekylet består af en lang kæde af såkaldte nukleotider, hvoraf der findes fire forskellige: A, C, G og U. Rækkefølgen af disse nukleotider bestemmer de genetiske egenskaber, og kaldes RNA'ets rygrad eller primærstruktur. I cellerne ligger RNA dog ikke bare som en lang kæde af nukleotider, for nukleotiderne er meget villige til at lave brintbindinger med andre nukleotider. Der er dog visse regler for, hvilke bindinger der kan opstå – de såkaldte Watson-Crick regler. A kan kun binde til U, mens G kan binde til C og U. Ethvert nukleotid kan kun binde til ét andet nukleotid. Et sådant par hedder et basepar. Baseparrene udgør sekundærstrukturen og bestemmer RNA'ets geometriske form. Der findes også tertiærstruktur, men det vil vi ignorere her.

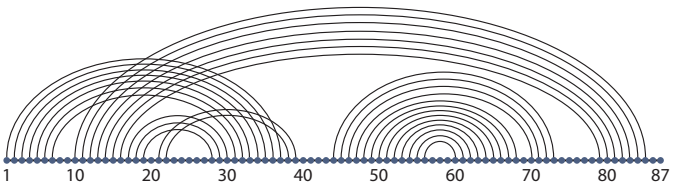
I naturen ser man, at to kæder med den samme rækkefølge af nukleotider (samme rygrad), folder til den samme geometriske form. Derfor burde det være muligt at forudse sekundærstrukturen alene ud fra rygraden. Det har man forsøgt i rigtig mange år. Der er udviklet mange opskrifter på, hvordan man finder frem til den rigtige struktur, en såkaldt algoritme, men ingen gætter rigtig godt. Den aarhusianske *gfold*-algoritme er nu den mest præcise og gætter 90 % af sekundærstrukturen rigtigt.

RNA som "fede grafer"

Et vigtigt redskab i at udvikle algoritmen er at se anderledes på RNA. I stedet for at se RNA



Figur 1a. En RNA-sekvens illustreret som et plant diagram.

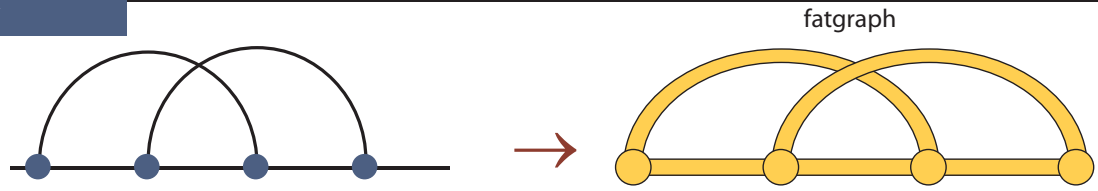


Figur 1b. Diagrammet viser samme RNA-sekvens som figur 1a.

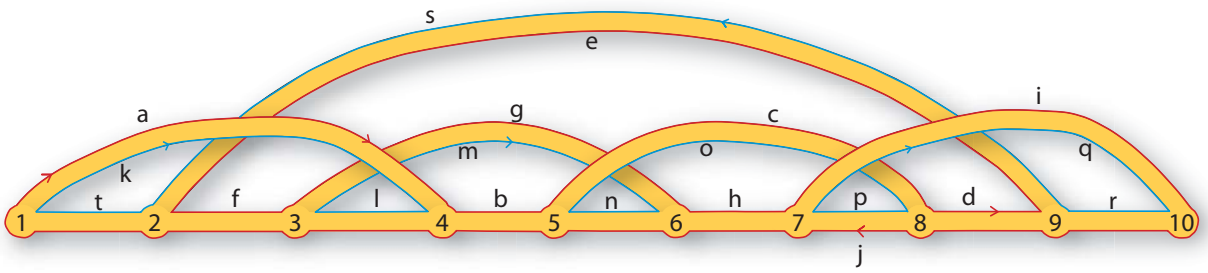
RNA består af en rygrad af nukleotider og bindinger mellem nukleotiderne. Vi vil se nukleotiderne som knudepunkter og rygraden som den sorte streg, der forbinder dem. De røde streger er brintbindingerne mellem nukleotiderne.

Ved at tage fat i begge ender af RNA'et og rette det ud, får vi en lige horisontal rygrad, og vi markerer de gamle brintbindinger med buer over rygraden (figur 1b). Rygraden ses som en horisontal linje med knudepunkter, og brintbindingerne som buer ovenover.

Fra RNA til fatgraph



Figur 2.



Figur 3.

For at få lavet en fatgraph er idéen at erstatte alle buerne med ikke-tvistede bånd, og alle knudepunkter (nukleotiderne) med små diske. Processen kan ses i figur 2, og på det højre billede ses de små diske og bånd, der er sat sammen. På den måde får vi et objekt, der har kanter.

Afhængig af, hvordan knudepunkterne er forbundet i den oprindelige graf, får vi et forskelligt antal sammenhængende kanter, ved simpelthen at følge en kant fra disk til disk, og fortsætte ad kanten, når man møder en disk (figur 3). Lad os lave et eksempel. Start i punktet 1 helt til venstre, og gå nu langs kanten på ydersiden af buen, hen til punkt 4. Herefter tager vi den første kant vi møder, som er oversiden af rygraden, der løber frem til punkt 5. Herefter tager vi igen oversiden af buen fra punkt 5 til 8. Sådan fortsætter vi, indtil vi kommer tilbage til udgangspunktet. Vi har nu fået den røde kant *abcedfghij*. For at finde alle kanter af fatgraphen, skal vi gennemløbe både inderside og yderside af hvert bånd, vi har opfedet buerne til. For at finde den sidste kant start nu igen ved punkt 1, og løb nu langs indersiden af båndet til punkt 4. Derefter løbes af oversiden af rygraden til punkt 3, og herefter på indersiden af båndet fra punkt 3 til punkt 6. Sådan fortsættes til vi kommer tilbage til punkt 1. Det giver den blå kant *klmnopqrst*. Hvis der er flere kanter, som mangler at blive gennemløbet, fortsættes på denne måde.

som en foldet streng (som i figur 1a), skal man i stedet tage fat i hver ende af RNA'et og strække det ud, så rygraden ligger som en horisontal linje, og brintbindingerne er repræsenteret af buer ovenover (se figur 1b). Herved får vi et diagram af RNA'et, der ser anderledes ud, men indeholder den samme information.

For at få matematik ind i billedet skal diagrammerne derefter konverteres til såkaldte fatgraphs. Det er nemlig fra fatgraphs, vi skal finde det værktøj, vi skal bruge. Der er en grund til, at det hedder fatgraphs. De laves fra diagrammer (el. grafer) ved at fede dem op.

Vi har nu taget et helt essentielt skridt: vi har oversat informationen, der er gemt i foldningen af RNA'et, til et matematisk objekt. Når RNA er udstillet som en fatgraph, kan vi hive værktøjskassen med matematik frem og bruge mange forskellige redskaber til at manipulere med RNA'et. I sidste ende kan vi gå tilbage til biologiens verden og lave forudsigelser, der kan testes i naturen.

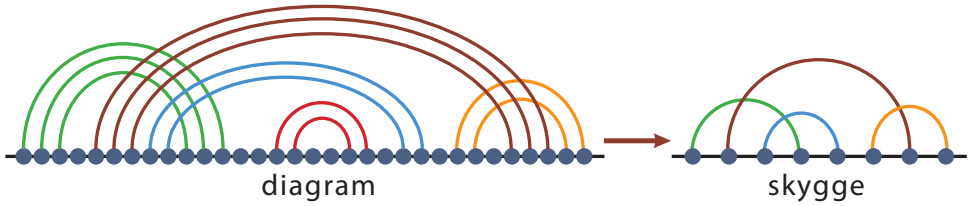
Genus af fatgraph

Genus kan beregnes direkte fra billedet af fatgraphen. Genus er defineret ved følgende formel:

$$g = \frac{1}{2}(1 + n - r),$$

hvor n er antallet af buer over rygraden (dvs. bindinger mellem nukleotiderne) og r er antallet af kanter. Genus er en såkaldt invariant, og kan beregnes og defineres på mange måder – ovenstående formel er blot en af mange måder.

Vi kan nemt se, at fjerner vi en af buerne i en stak af buer, reducerer vi n med 1 men samtidig bliver r også 1 mindre, så g er uændret. Det er derfor, at vi reducerer alt til skygger.



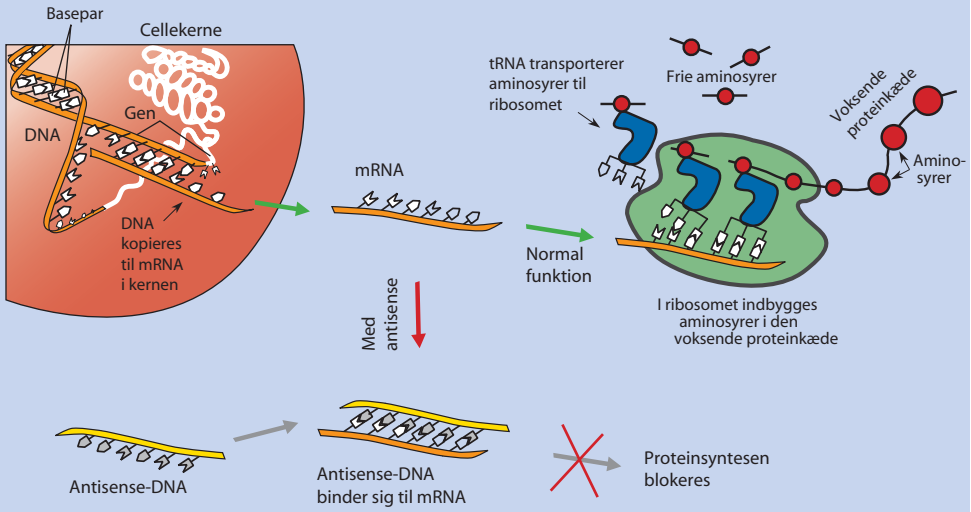
Figur 4. Reduktion af diagram til en skygge. Skyggen indeholder præcis nok information til, at kunne bestemme genus. Derfor fjernes knudepunkter, der ikke er forbundet til andre via brintbindinger. Da genus ikke afhænger af antallet af buer i hver af stak, men kun af antallet af stakke, kollapses stakkene til en enkelt bue.

Skygger af den fede graf

Det essentielle matematiske værktøj, vi skal bruge, er et tal, der kaldes en fatgraphs *genus*. Genus af fatgraphen er nøglen til at få algoritmen til at virke.

Genus finder vi ved at reducere diagrammerne, så kun den essentielle information er tilbage. Når brintbindingerne i RNA er angivet som buer ovenover en rygrad, er det kun de buer, der krydser hinanden, der bridrager til genus. Derfor fjerner vi alle de buer, der ikke krydser hinanden, vi fjerner alle de knudepunkter, der ikke har en bue tilknyttet, og vi erstatter alle stakke af buer med en enkelt bue (se figur 4). Herved får vi en struktur, vi

Antisense slukker gener



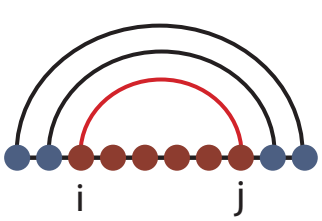
Antisense-princippet går ud på at "slukke" for et givent gen, der er årsag til en sygdom. Et gen består af et stykke DNA med en bestemt sekvens af nukleotider. Når et gen aktiveres, dannes der en kopi af dette stykke DNA – såkaldt *messenger RNA* (mRNA). Denne kopi transporteres ud af cellekernen til et organel i cellen kaldet ribosomet. Her "oversættes" informationen i RNA-molekylet til et protein, som så udfører en given funktion i organismen. Hvis genet er "sygt" kan funktionen af proteinet være skadelig for organismen.

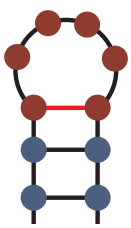
Mekanismen i antisense er, at man konstruerer en RNA-streng, der består af en sekvens af nukleotider, der præcis modsvarer sekvensen i det gen, man gerne vil have sat ud af spillet. Denne streng vil så binde sig til genets budbringer – mRNA'et – som derved bliver dobbeltstrenget. Det betyder, at informationen ikke bliver oversat til protein, og genet er dermed blevet "slukket".

RNA er biologisk aktivt ved nukleotider, der ikke indgår i basepar med andre nukleotider. Sammenlimningen af den kunstigt fremstillede RNA-sekvens og RNA'et, sker ved de biologiske aktive områder. Det er derfor essentielt at vide, hvor disse områder er. Man skal med andre ord kende RNA'ets sekundærstruktur, og det er netop denne, som den århusianske *gfold*-algoritme kan udregne.

Det første medicinske præparat, der virker ved antisense, kom på markedet i 1998 i USA. Præparatet modvirker en virus, der giver inflammation på nethinden i øjet, og kan gøre en blind. I de seneste år har der været et øget fokus på medicin, der virker ved antisense, bl.a. i forbindelse med medicin mod HIV.

Figur 5: Disse fire skygger er de eneste sammenhængende skygger, der giver fatgraphs af genus 1.





Figur 6. Et såkaldt Hairpin-loop, som er en biologisk aktiv struktur, der bl.a. bruges til at binde til andre molekyler. Til højre ses den geometriske udformning, og til venstre det tilsvarende diagram. Den røde bue viser skyggen. Skyggen har genus 0.

kalder skyggen af diagrammet.

Et diagram kaldes sammenhængende, hvis alle buer er forbundet til hinanden ved hjælp af andre buer. En skygge er ikke nødvendigvis sammenhængende, men er sammensat af blokke af sammenhængende delskygger. Det smarte ved denne opdeling af skyggerne er, at genus af en skygge er summen af genus af de sammenhængende delskygger. Et vigtigt skridt i algoritmen er at bestemme antallet af sammenhængende skygger af en vis genus.

Kræver enorm regnekraft

For at forudsige RNA-foldningerne, skal der laves mange beregninger. Det kræver så megen regnekraft, at der selv med de største computercentre ikke er kapacitet til at kunne forudsige foldningen af et helt RNA-molekyle. Derfor er det nødvendigt at begrænse sig til korte RNA-sekvenser – dvs. RNA-strenge med en rygrad bestående af forholdsvis få nukleotider. Derudover begrænser man sig også i, hvor komplekse strukturer ens algoritme skal medtage. Hvis strukturerne stiger i kompleksitet, vokser kravet om regnetid meget hurtigt. Derfor har tidligere algoritmer kun beskæftiget sig med skygger af genus 0. Det er dog en uheldig begrænsning, da man har observeret RNA i naturen, der inkluderer krydsende buer, og genus af en sådan RNA-streng er dermed større end 0.

gfold-algoritmen inkluderer netop krydsende buer. Men for at undgå, at problemet bliver for komplekst, nøjes vi med at se på RNA-sekvenser, der har en underliggende skygge, opbygget af sammenhængende delskygger, som hver har genus 1. Altså kan algoritmen håndtere alt som tidligere har været betragtet af biologerne, samtidig med en del nye foldninger.

I den matematiske disciplin kombinatorik har man i mange år set på disse skyggediagrammer, uden man kendte til koblingen til RNA. Kombinatorik-

kerne har beregnet, at i genus g findes der kun endeligt mange skygger. Det er helt afgørende, for hvis man vil lave en algoritme, hvor man prøver sig systematisk frem, er det vigtigt, at der kun er endeligt mange kombinationsmuligheder – ellers bliver algoritmen aldrig færdig.

Den aarhusianske algoritme kan som nævnt håndtere RNA-sekvenser opbygget af skygger med højest genus 1, og dem findes der kun fire forskellige af (de er vist i Figur 5).

Beregning i to trin

gfold-algoritmen består essentielt af to trin. Første trin er at finde frem til alle de mulige kombinationer af sammenhængende skygger med genus 1, som findes på en given RNA-sekvens. Andet trin er at udvælge den rigtige af de mange foreslåede RNA-sekvenser. Og her skal vi bruge fysiske overvejelser. Naturen er sådan indrettet, at den altid søger mod tilstande med mindst mulig energi. Det kan f.eks. forklare, hvorfor det Aktuel Naturvidenskab, du holder i hånden, falder til jorden, hvis du slipper det. Energien af bladet er lavest ved jorden. Der findes en måde, hvorpå energien af en RNA-struktur kan beregnes. Energien afhænger naturligvis af, hvilke nukleotider der bliver parret, antallet af bindinger i hver stak af bindinger, og flere andre parametre. Pointen er dog, at denne energi let kan beregnes fra en foreslået RNA-struktur. Da naturen altid vælger den løsning med lavest energi, er det nu let at finde den rigtige RNA-struktur fra de endeligt mange som algoritmens første del har produceret.

Ikke i mål endnu

Selvom gfold-algoritmen tydeligt forbedrer de eksisterende algoritmer, er målet langtfra nået. Algoritmen inddrager mere kompleksitet end tidligere algoritmer, men det er desværre – i hvert fald indtil videre – på bekostning af en væsentligt højere regnetid end de bedste konkurrerende algoritmer. Det betyder, at hidtidige simuleringer er lavet på RNA-sekvenser med 150 nukleotider – hvilket for en anvendelsesorienteret betragtning ikke er brugbart. Til sammenligning indeholder et helt menneskeligt RNA 3 milliarder nukleotider. Det er dog ikke nødvendigt at forudsige foldningen af hele RNA'et. For at kunne lave Antisense er det nok at vide, hvordan det enkelte gen folder. Det gennemsnitlige antal basepar i et gen er 3.000, men det varierer dog meget fra gen til gen.

Den største udfordring i dag er derfor at nedbringe regnetiden. Det kræver, at algoritmen optimeres. Det arbejdes der stadig på, i tæt samarbejde med forskere fra Kina og Tyskland. De næste år bliver derfor meget interessante for både matematikere og biologer. ■

Forfatteroplysninger

Jakob Blaavand er ph.d.-studerende ved QGM, Aarhus Universitet og University of Oxford.
blaavand@qgm.au.dk

Læs mere

Reidys, C. et. al. "Topology and prediction of RNA pseudoknots", Bioinformatics vol 27 issue 8, 2011.
Til ovenstående findes "Supplementary material", der giver mange eksempler.

Mere om antisense:
Lichtenstein, C. og Nellen, W. "Antisense Technology: A Practical Approach" bog udgivet på Oxford University Press, 1998.