Matthias Mnich
Faculteit Wiskunde & Informatica
Technische Universiteit Eindhoven
Postbus 513
5600 MB Eindhoven
m.mnich@tue.nl

Onderzoek

Allemaal op een rijtje

De laatste decennia is er veel onderzoek gedaan naar het menselijk genoom. En met succes. Inmiddels is de volledige genetische code van de mens ontrafeld. Wetenschappers hopen hiermee te ontdekken welke genen verantwoordelijk zijn voor welke ziektes. Zij kijken daarbij naar zogenaamde markers op het chromosoom, en zijn met name geïnteresseerd in de volgorde waarin die markers voorkomen. Matthias Mnich van de Technische Universiteit Eindhoven bedacht een nieuwe, snellere methode om die volgorde vast te stellen. Met zijn voordracht over dit onderzoek won Matthias Mnich in 2010 de Philips Wiskundeprijs voor Promovendi.

Het Human Genome Project was een internationaal wetenschappelijk onderzoeksproject met als eerste doel om voor elk chromosoom de volgorde (sequentie genoemd) van de nucleotiden, de bouwstenen van ons DNA, vast te stellen, en daarnaast om de 20.000 tot 25.000 genen in het menselijk genoom te identificeren en in kaart te brengen, zowel qua locatie als qua functie. Het project startte in 1990; een voorlopige versie van het genoom werd gepubliceerd in 2000 en een vrijwel complete versie in 2003, maar nog steeds verschijnen er nieuwe resultaten. Het in kaart brengen van menselijke genen is een belangrijke stap in de ontwikkeling van medicijnen en speelt een belangrijke rol in andere gebieden van de gezondheidszorg. Het genoom van een willekeurig individu (op een-eïge tweelingen en gekloonde organismen na) is uniek: het in kaart brengen van het menselijk genoom vereist dus het bestuderen van verschillende variaties van elk gen. Niet het hele DNA van menselijke cellen is tot nu toe bekeken; ongeveer acht procent van het totaal is nog niet bestudeerd. Een reden hiervoor is het feit dat we daarbij tegen grenzen aanlopen, die ons voor grote wetenschappelijke uitdagingen stellen. Wiskunde, en combinatoriek in het bijzonder, kunnen een grote rol spelen om hier uitkomst in te bieden.

Een van de problemen waar biologen tegen aanlopen is het volgende. Een *marker* is een deelrijtje (of gen) van het dubbele DNA-molecuul dat precies één keer voorkomt in het menselijke genoom. Om de werking van markers te begrijpen, zijn biologen geïnteresseerd in hun volgorde op het chromosoom. Vanwege de methode waarop de data worden verkregen, kunnen we alleen voor drietallen markers de volgorde vinden. Bovendien weten we voor die drietallen alleen maar welke van de drie markers in het midden staat: $b < a < c$ en $c < a < b$ zijn dus ononderscheidbaar. De input bestaat dus uit paren van de vorm $(a, \{b, c\})$, die staan voor de bewering dat "marker a tussen markers b en c ligt". Precies de twee lineaire ordeningen $b < a < c$ en $c < a < b$ *voldoen aan* de voorwaarde $(a, \{b, c\})$. Gegeven zo'n collectie $\mathcal{C}$ van paren van deze vorm, is ons doel een lineaire ordening van de markers te vinden die voldoet aan een zo groot mogelijk aantal voorwaarden in $\mathcal{C}$.

Een triviale manier om dit doel te bereiken, is om *alle* lineaire ordeningen van markers te bekijken en voor elk van hen te tellen aan hoeveel voorwaarden in $\mathcal{C}$ ze voldoen. Voor n markers gaat dit om $n!/2$ lineaire ordeningen. Echter, voor $n = 60$ overschrijdt de waarde $n!/2$ al 10^{80}, wat ongeveer het aantal atomen in het waarneembare heelal is. Aangezien biologen te maken hebben met $n \approx 3.000 - 5.000$ markers tegelijk, moeten wij als wiskundigen behoorlijke inspanningen doen om nog serieus genomen te worden. In de literatuur zijn enkele heuristische oplossingen en benaderingen gesuggereerd voor dit probleem,

Figuur 1 Een deel van een DNA-molecuul met daarop aangegeven een marker. Het DNA-molecuul kan bestaan uit tienduizend tot een miljard nucleotiden op een rij.

maar het op een niet-triviale manier vinden van een optimale lineaire ordening was nog steeds een open probleem. In het vervolg zal ik een eenvoudige en intuïtieve oplossing voor dit probleem beschrijven, waarin de aloude wiskundige structuur van lineaire ordeningen interessante dwarsverbanden laat zien tussen verschillende wiskundige vakgebieden en wellicht tot antwoorden leidt die ons het menselijk genoom beter doen begrijpen.

Optimale lineaire ordening

De ingrediënten van onze aanpak om optimale lineaire ordeningen te vinden zijn eenvoudige kansrekening, een beetje algebra en een computer. We zijn op zoek naar een lineaire ordening α van een marker set M die voldoet aan een maximaal aantal voorwaarden van de vorm "a ligt tussen b en c" uit een gegeven collectie $\mathcal{C}$.

Om zo'n lineaire ordening α te vinden, kijken we eerst, voor willekeurige collectie voorwaarden $\mathcal{C}$, naar het minimale deel van de voorwaarden uit $\mathcal{C}$ waaraan kan worden voldaan door een geschikte lineaire ordening. Voor elke voorwaarde uit $\mathcal{C}$ geldt wegens symmetrie dat precies een derde deel van de $n!/2$ mogelijke lineaire ordeningen eraan voldoet. Omgekeerd moet er dus een lineaire ordening zijn die voldoet aan minstens een derde deel van de voorwaarden uit $\mathcal{C}$. Beter dan deze $\frac{1}{3}$ kunnen we niet komen, vanwege de complete collecties $\mathcal{C}$ die bestaan uit de voorwaarden

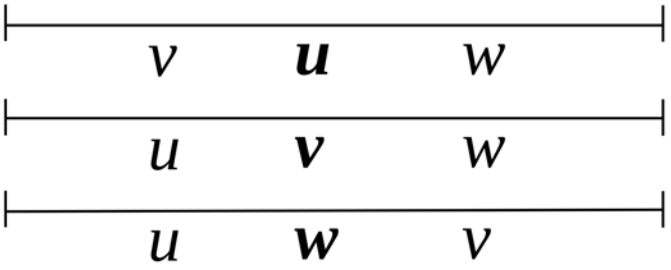
$$\left\{\begin{array}{l} C_1 = \text{“ } a \text{ ligt tussen } b \text{ en } c \text{ ”}, \\ C_2 = \text{“ } b \text{ ligt tussen } a \text{ en } c \text{ ”}, \\ C_3 = \text{“ } c \text{ ligt tussen } a \text{ en } b \text{ ”}. \end{array}\right\}$$

Elke lineaire ordening van M voldoet immers aan precies één van deze voorwaarden. In dit geval zijn de voorwaarden C_1, C_2 en C_3 in 'conflict' met elkaar, en zulke situaties kunnen voorkomen als biologen niet helemaal zeker zijn van de volgorde van a, b en c.

Dit brengt ons bij de volgende vereenvoudigingsregel:

Verwijder elk compleet drietal voorwaarden uit $\mathcal{C}$. (R)

Noem $\mathcal{C}'$ de overgebleven collectie van voorwaarden, M' de verzameling markers in $\mathcal{C}'$, en r het aantal complete drietallen voorwaarden dat door regel (R) verwijderd is. Merk op dat na toepassing van stap (R) elk van de markers a, b en c nog steeds in M' kan zitten, maar ze komen in geen enkele voorwaarde uit $\mathcal{C}'$ meer alle drie tegelijk voor. Nu bevat $\mathcal{C}'$ precies $3r$ voorwaarden minder dan $\mathcal{C}$, waarvan er voor elke lineaire ordening van M precies r zijn waaraan deze ordening voldoet. We concluderen dat het lukt om aan minstens $|\mathcal{C}|/3 + k$ voorwaarden uit $\mathcal{C}$ te voldoen precies dan als er aan minstens $|\mathcal{C}'|/3 + k$ voorwaarden uit $\mathcal{C}'$ kan worden voldaan. Kortom, toepassen van regel (R) leidt tot een equivalent probleem en we mogen er in het vervolg van uitgaan dat onze collectie voorwaarden $\mathcal{C}$ geen complete drietallen bevat.



Figuur 2 Drie voorwaarden op dezelfde drie markers a, b en c, waar de bovenste figuur de voorwaarde "a ligt tussen b en c" voorstelt. Elke lineaire ordening van a, b, c voldoet precies aan een van deze drie voorwaarden.

Zou het lukken om aan meer voorwaarden te voldoen in een dergelijke *gereduceerde* collectie voorwaarden $\mathcal{C}$? Oftewel, kunnen we voor elke niet-triviale gereduceerde collectie $\mathcal{C}$ aan minstens $|\mathcal{C}|/3 + h(|\mathcal{C}|)$ voorwaarden voldoen, voor zekere positieve waarde $h(|\mathcal{C}|)$? Als we deze vraag bevestigend kunnen beantwoorden, hebben we meteen een optimale lineaire ordening α voor de oorspronkelijke verzameling markers M, door eerst $\mathcal{C}$ met behulp van (R) te reduceren, dan een optimale ordening α' te vinden van M', en ten slotte α' weer willekeurig uit te breiden tot een lineaire ordening α van M. Hoe groter $h(|\mathcal{C}|)$ is, des te minder markers M' zullen er blijken over te blijven, dus des te sneller kan α' worden gevonden.

We zullen nu aantonen dat er inderdaad zo'n positieve waarde $h(|\mathcal{C}|)$ bestaat, en wel van orde $\sqrt{|\mathcal{C}|}$. Om $h(|\mathcal{C}|)$ te bepalen, beschouwen we wederom een random lineaire ordening π van M, maar nu eentje die in twee stappen verkregen is. Deze methode lijkt misschien op het eerste gezicht onnodig ingewikkeld vergeleken met gewoon direct een willekeurige random ordening op M kiezen (uit de $n!/2$ mogelijkheden), maar verderop zal deze methode nuttig blijken.

De eerste stap is om de verzameling markers M op willekeurige wijze over vier vazen te verdelen, genummerd 0, 1, 2 en 3, waarbij het aantal markers per vaas dus kan variëren. Voor elke marker a uit M schrijven we $\phi(a)$ voor het vaasnummer waarin marker a is terecht gekomen. In de tweede stap verfijnen we deze partitie ϕ van M, door eerst de markers uit vaas 0 in een random volgorde te leggen, vervolgens daarachter de markers uit vaas 1, et cetera. Gegeven een voorwaarde C en een partitie ϕ, bekijken we voor alle lineaire ordeningen π die behoren bij ϕ de verwachtingswaarde $X_C = \mathbb{E}[\chi_C(\pi)] - 1/3$ van de indicatorfunctie $\chi_C(\pi)$ die 1 (dan wel 0) is als π voldoet aan C (of niet). Door de partitie ϕ random te kiezen, verkrijgen we op deze manier een stochast X_C. De reden om X_C op deze manier te definiëren, is dat we voor $X = \sum_{C\in\mathcal{C}} X_C$ nu weten: als $X \geq k$, dan wordt aan ten minste $|\mathcal{C}|/3 + k$ voorwaarden uit $\mathcal{C}$ voldaan. Het is nu dus voldoende om een positieve waarde van $h(|\mathcal{C}|)$ te vinden, zodanig dat $X \geq h(|C|)$ geldt met positieve kans voor elke gereduceerde niet-triviale collectie $\mathcal{C}$.

Invloed van de momenten

Om zo'n positieve waarde van $h(|\mathcal{C}|)$ te vinden, bekijken we de momenten van X. We beginnen met de verwachtingswaarde $\mathbb{E}[X]$ van X. Voor elke voorwaarde C = "a is between b and c", bekijk de vazen waartoe a, b en c behoren. De kans dat alle drie deze markers in dezelfde vaas zitten, dat wil zeggen, dat $\phi(a) = \phi(b) = \phi(c)$, is gelijk aan $1/16$, en in dit geval is de waarde van X_C gelijk aan nul. We laten het bestuderen van X_C voor de andere verdelingen van a, b en c over de vier vazen aan de lezer over. We vinden dat $\mathbb{E}[X_C] = 0$, en dus wegens lineairiteit van de verwachtingswaarde ook $\mathbb{E}[X] = 0$.

We bekijken nu de variantie $\mathbb{E}[X^2]$ van X. Vergeleken met $\mathbb{E}[X]$ is dit een stuk meer werk. In feite zullen we slechts een ondergrens voor $\mathbb{E}[X^2]$ vinden en zelfs hiervoor hebben we de hulp van een computer nodig. We geven nu een schets van deze aanpak.

Door haakjes uit te werken, splitsen we de berekening van $\mathbb{E}[X^2] = \mathbb{E}[(\sum_{C\in\mathcal{C}} X_C)^2]$ als volgt op:

$$\mathbb{E}[X^2] = \sum_{C\in\mathcal{C}} \mathbb{E}[X_C^2] + \sum_{\substack{(C,C')\in\mathcal{C}\times\mathcal{C} \\ C\neq C'}} \mathbb{E}[X_C X_{C'}]. \qquad (1)$$

Voor de kansverdeling van X, waarvoor we eerder vonden dat $\mathbb{E}[X] = 0$, vinden we op analoge wijze dat $\mathbb{E}[X_C^2] = 11/96 = 88/768$. We tonen nu

aan dat

$$\sum_{(C,C')\in\mathcal{C}\times\mathcal{C},C\neq C'} \mathbb{E}[X_C X_{C'}] \geq -\frac{77}{768}|\mathcal{C}|.$$

Beschouw twee verschillende voorwaarden C en C'. Als C en C' geen markers gemeen hebben, dan zijn X_C en X'_C onafhankelijke stochasten, en dus geldt dan dat

$$\mathbb{E}[X_C X_{C'}] = \mathbb{E}[X_C]\mathbb{E}[X_C] = 0 \cdot 0 = 0.$$

Als C en C' wel een of meer markers gemeen hebben, dan zijn er een groot aantal gevallen te onderscheiden bij het berekenen van $\mathbb{E}[X_C X_{C'}]$. Maar met enkele pagina's rekenwerk en de hulp van een computer, komen we hier uiteindelijk wel uit. Bij deze analyse gebruiken we uiteraard het feit dat als C en C' voorwaarden zijn op dezelfde drie markers, dan de derde voorwaarde op die markers niet in de collectie voorwaarden $\mathcal{C}'$ voorkomt. We concluderen dat

$$\mathbb{E}[X^2] \geq 11/768|\mathcal{C}|.$$

Het derde moment van X is voor ons niet van belang, maar het vierde moment $\mathbb{E}[X^4]$ daarentegen wel. Het berekenen van $\mathbb{E}[X^4]$ op dezelfde manier als $\mathbb{E}[X^2]$, is echter onbegonnen werk. Maar we kunnen wel onze toevlucht zoeken tot de zogenaamde *Hypercontractive Inequality*, een standaard gereedschap voor de functionaalanalyticus. Deze Hypercontractive Inequality geeft ons bovengrenzen voor het vierde moment van een stochast X in termen van de variantie van X, zolang de waarden van X gelijk zijn aan de waarden van een meerdimensionaal polynoom van vaste graad in een random vector met componenten -1 en $+1$. Om precies te zijn, als $X = f(\epsilon_1, \ldots, \epsilon_n)$ voor een reëelwaardig polynoom f van graad r in n variabelen en een random vector $(\epsilon_1, \ldots, \epsilon_n) \in \{-1, +1\}^n$, dan geldt dat $\mathbb{E}[X^4] \leq 9^r \mathbb{E}[X^2]^2$. Deze bovengrens voor $\mathbb{E}[X^4]$ zal goed genoeg blijken voor ons doel. Op het eerste gezicht lijkt er niet een intuïtieve keuze te zijn voor het polynoom f van vaste graad, omdat X ons informatie geeft over lineaire ordeningen waarin alle markers maar liefst n posities kunnen innemen. Hier blijkt de partitie ϕ echter weer uitkomst te bieden.

Voor elke voorwaarde $C =$ "a ligt tussen b en c" definiëren we de variabelen

$$\epsilon_a^1 = \begin{cases} -1, & \phi(a) \in \{0,1\}, \\ +1, & \phi(a) \in \{2,3\}, \end{cases} \qquad \epsilon_a^2 = \begin{cases} -1, & \phi(a) \in \{0,2\}, \\ +1, & \phi(a) \in \{1,3\}, \end{cases}$$

en analoog $\epsilon_b^1, \epsilon_b^2$ en $\epsilon_c^1, \epsilon_c^2$. Dan kan $\epsilon_a^1\epsilon_a^2$ gezien worden als de binaire representatie van een getal uit $\{0,1,2,3\}$ en $\epsilon_a^1\epsilon_a^2\epsilon_b^1\epsilon_b^2\epsilon_c^1\epsilon_c^2$ kan gezien worden als de binaire representatie van een getal uit $\{0,1,\ldots,63\}$, waarbij -1 de rol van 0 vertolkt. Tot groot genoegen van de algebraïcus, kunnen we de stochast X_C uitdrukken als het polynoom

$$X_C = \frac{1}{64}\sum_{q=0}^{63}(-1)^{s_q} w_q \cdot (\epsilon_a^1 + v_{qa}^1)(\epsilon_a^2 + v_{qa}^2)(\epsilon_b^1 + v_{qb}^1) \cdot (\epsilon_b^2 + v_{qb}^2)(\epsilon_c^1 + v_{qc}^1)(\epsilon_c^2 + v_{qc}^2)$$

waarbij $v_{aq}^1 v_{aq}^2 v_{bq}^1 v_{bq}^2 v_{cq}^1 v_{cq}^2$ de binaire representatie is van q, s_q het aantal cijfers gelijk aan -1 in die representatie en w_q de waarde van X_C wanneer de binaire respresentaties van $\phi(a)$, $\phi(b)$ en $\phi(c)$ gelijk zijn aan respectievelijk $v_{aq}^1 v_{aq}^2$, $v_{bq}^1 v_{bq}^2$ en $v_{cq}^1 v_{cq}^2$. Dit polynoom is van de graad $r = 6$ en dus is ook $X = \sum_{C\in\mathcal{C}} X_C$ uit te drukken als een zesdegraads polynoom. Wegens de Hypercontractive Inequality geldt nu dus dat $\mathbb{E}[X^4] \leq 9^6\mathbb{E}[X^2]^2$.

Op dit moment weten we over X dat $\mathbb{E}[X] = 0$, $\mathbb{E}[X^2] = \sigma^2 > 0$ en $\mathbb{E}[X^4] \leq c\sigma^4$, voor zekere constante c. Dus neemt X volgens een resultaat van Alon et al. [2] met positieve kans een waarde groter dan $\sigma/2\sqrt{c}$ aan. Invullen van $c = 9^6$ en de ondergrens voor σ geeft nu het volgende resultaat: van elke gereduceerde collectie voorwaarden $\mathcal{C}$ kan aan ten minste

$$\frac{|\mathcal{C}|}{3} + h(|\mathcal{C}|) = \frac{|\mathcal{C}|}{3} + \frac{1}{46656}\sqrt{\frac{11}{3}|\mathcal{C}|}$$

voorwaarden worden voldaan.

Laten we nog concreter bekijken hoe we een optimale lineaire ordening α van alle markers M die voorkomen in de collectie voorwaarden $\mathcal{C}$ kunnen verkrijgen. Ten eerste, gegeven de collectie $\mathcal{C}$, reduceren we $\mathcal{C}$ door middel van regel (R) tot een collectie $\mathcal{C}'$ met verzameling markers M'. Dan kunnen we achtereenvolgens voor elke $k = 2/3|\mathcal{C}'|, 2/3|\mathcal{C}'| - 1, \ldots, 0$ onderzoeken of er aan $|\mathcal{C}'|/3 + k$ voorwaarden uit $\mathcal{C}'$ te voldoen is en we stoppen zodra we zo'n k hebben gevonden. Merk op dat

$$k \leq \frac{1}{46656}\sqrt{\frac{11}{3}|\mathcal{C}'|}$$

altijd voldoet, terwijl

$$k > \frac{1}{46656}\sqrt{\frac{11}{3}|\mathcal{C}'|}$$

betekent dat we door het toepassen van (R) de collectie $\mathcal{C}$ hebben 'samengeperst' tot slechts $O(k^2)$ voorwaarden. Zodra we ten slotte α' gevonden hebben, kunnen we deze gemakkelijk uitbreiden tot een optimale volgorde α van M.

Andere ordeningsproblemen

Tot nu toe hebben we *betweenness* bestudeerd, met voorwaarden van de vorm $C =$ "a ligt tussen b en c". Maar we zouden ook kunnen denken aan voorwaarden van de vorm $C = (a,b,c)$, waarmee we de precieze volgorde van de markers a, b en c op de lijn bedoelen. Verder hebben we de markers steeds geordend op een lijn, maar ordeningen op andere meetkundige objecten zijn ook interessant. Echter, als drie variabelen a, b en c op een cirkel staan, dan staat elk van de drie 'tussen' de andere twee, dus moeten we nu naar equivalentieklassen van ordeningen kijken. Meer in het algemeen krijgen we een hele klasse aan ordeningsproblemen, één voor elke deelverzameling S van de permutatiegroep S_3 die uit alle permutaties van 3 elementen bestaat, waarbij S vastlegt wat de toegestane permutaties zijn voor een voorwaarde opdat aan die voorwaarde wordt voldaan. Betweenness komt bijvoorbeeld overeen met $S = \{(123), (321)\}$. Uiteraard, als alle permutaties toegestaan zijn, of juist geen enkele, dus als $S = S_3$ of $S = \emptyset$, dan is het ordeningsprobleem triviaal. Het lijkt erop dat er voor alle overige $2^{|S_3|} - 2 = 62$ keuzes voor S geen efficiënte algoritmes voor het ordeningsprobleem bestaan. Gelukkig is onze probabilistische aanpak uit te breiden van betweenness tot de hele klasse van ordeningsproblemen, waarmee we goede combinatorische grenzen voor elk van hen kunnen bewijzen. Daarvoor moeten we echter nog wel wat doen: we

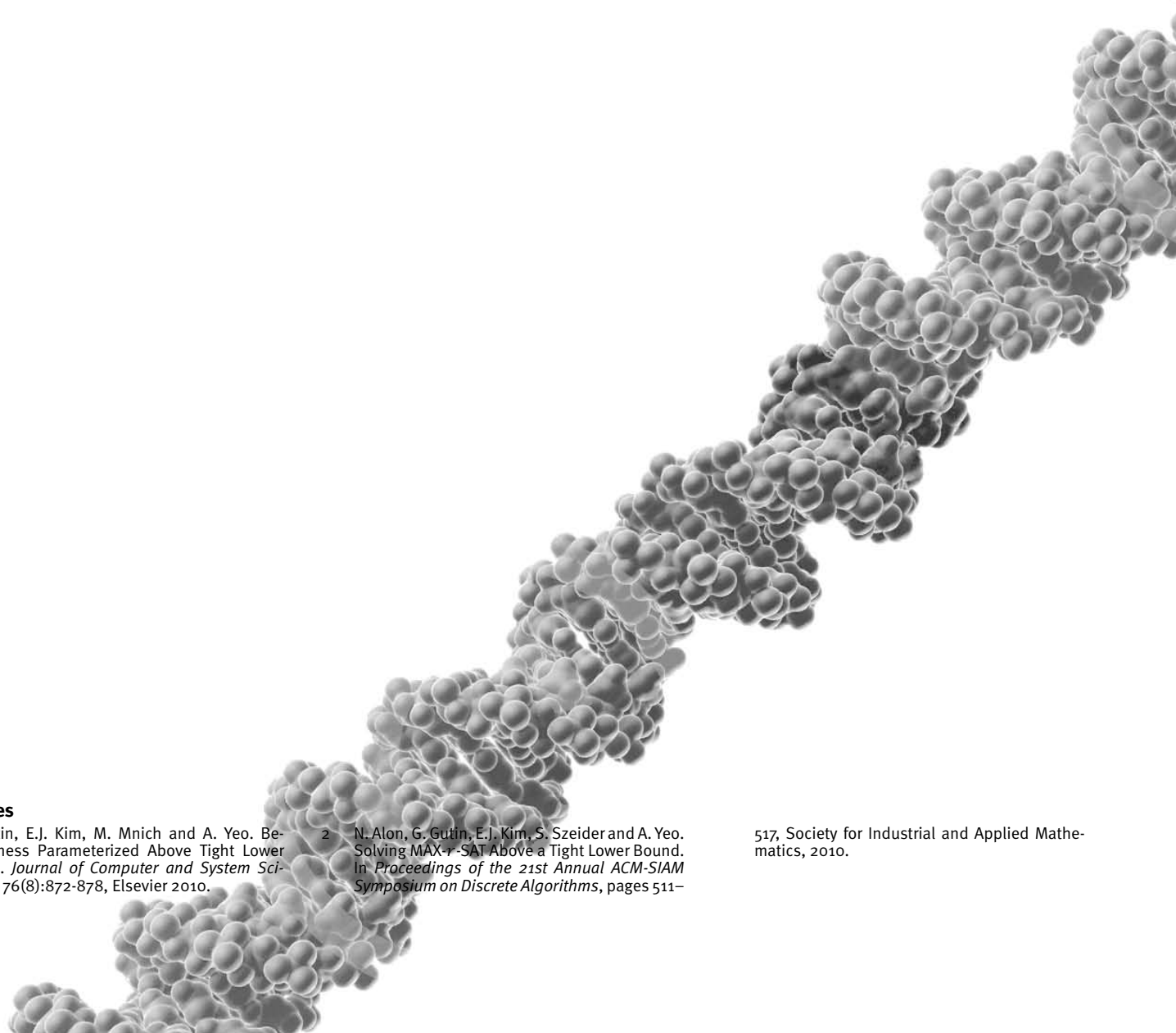
kunnen bijvoorbeeld aantonen dat we — in geval van $S = \{(123)\}$ — *oneindig* veel simplificatieregels nodig hebben, terwijl we in het geval van betweenness aan de enkele regel (R) genoeg hadden.

We hebben gezien dat het aloude concept van lineaire ordeningen nog steeds leidt tot interessante nieuwe onderzoeksvragen. Verbanden met andere vakgebieden binnen de wiskunde komen aan het licht, net als toepassingen die betrekking hebben op fundamenteel begrip van het leven. De resultaten die we tot nu toe verkregen hebben zijn nog verre van direct toepasbaar. De probabilistische aanpak is eenvoudig, maar voor practische toepassing in de biologie zou $h(|\mathcal{C}|)$ nog veel hoger moeten zijn.

Een andere interessante onderzoeksrichting is ordeningsproblemen met voorwaarden op meer dan drie markers. Voorwaarden op meer markers, zowel als ordeningen op cirkels, spelen een belangrijke rol in ruimtelijk redeneren — een ander hoofdstuk in het begrijpen van de *order of life*.

Dankwoord

Het onderzoek beschreven in dit artikel is gebaseerd op gezamenlijk werk met Gregory Gutin, Eun Jung Kim en Anders Yeo van de Royal Holloway - University of London, en verschijnt onder de titel 'Betweenness Parameterized Above Tight Lower Bound' in het *Journal of Computer and System Sciences*. We bedanken NWO voor hun steun, grant 639.033.403. De auteur bedankt Quintijn Puite voor zijn hulp bij het maken van de Nederlandse versie van dit artikel.



Referenties

1 G. Gutin, E.J. Kim, M. Mnich and A. Yeo. Betweenness Parameterized Above Tight Lower Bound. *Journal of Computer and System Sciences*, 76(8):872-878, Elsevier 2010.

2 N. Alon, G. Gutin, E.J. Kim, S. Szeider and A. Yeo. Solving MAX-r -SAT Above a Tight Lower Bound. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 511–517, Society for Industrial and Applied Mathematics, 2010.