PORAVNAVA NIZOV IN DELANNOYJEVA ŠTEVILA

MARKO RAZPET

Pedagoška fakulteta v Ljubljani

Math. Subj. Class. (2010): 05A15, 40B05, 68R15, 92D20

Pokazali bomo, kako je poravnava nizov povezana z Delannoyjevimi števili $D(m,n)$, s katerimi zapišemo število mrežnih poti v množici $\mathbb{N}\times\mathbb{N}$ od točke $(0,0)$ do dane točke (m,n), pri čemer so dovoljeni koraki v smereh $(1,0)$, $(0,1)$ in $(1,1)$. Vpeljali bomo Levenštejnovo razdaljo med nizoma in na kratko predstavili njeno uporabo v genetiki.

SEQUENCE ALIGNMENT AND DELANNOY NUMBERS

We will show how the sequence alignment is connected with the Delannoy numbers $D(m,n)$ which express the number of lattice paths in the set $\mathbb{N}\times\mathbb{N}$ with allowed steps $(1,0)$, $(0,1)$ and $(1,1)$. The Levenshtein distance between two sequences is introduced and its application in genetics is briefly presented.

Uvod

Kdor piše z računalniškim urejevalnikom besedil, zagotovo pozna poravnavo vrstic: levo, desno, sredinsko in obojestransko. Po navadi imamo najraje obojestransko poravnano besedilo, pri čemer sta vnaprej določena levi in desni rob, ki nam kot vidna ali nevidna ravna črta ne dovoljujeta, da bi kakšna črka, številka, ločilo ali kakšen drug znak padel bolj v levo od levega roba oziroma bolj v desno od desnega roba. Pomembno vlogo pri tem igra tudi znak za presledek, kateremu je namenjena posebna tipka. S presledki na primer ločujemo besede in nakažemo nov odstavek, lahko pa si mislimo, da z njimi na desno poravnamo prekratko vrstico. Ker običajno prostori za grafične znake niso enake širine, del težav s poravnavo odpade, ker urejevalniki po potrebi sami rahlo zgoščajo in redčijo razmike med znaki.

Če pa izberemo znake, ki zahtevajo enake širine, potem imamo z obojestransko poravnavo take težave kot nekoč na klasičnih pisalnih strojih. In ravno taka poravnava nas bo v nadaljevanju najbolj zanimala. Spoznali bomo, da je poravnava besed oziroma nizov kar uporabna. Vpeljemo namreč lahko mero, koliko sta si niza blizu. Ker pa so znaki ali elementi niza lahko marsikaj, lahko preverjamo ne le podobnost besedil in s tem odkrivamo na primer plagiatorstvo, ampak tudi podobnost tonskih zapisov in nukleotidnih zaporedij v genetiki. Nekaj več o tem bomo povedali v zadnjem razdelku.

Poravnava nizov

Prej povedano bomo sedaj posplošili. Iz črk, številk, ločil in morebitnih drugih znakov nekega področja, ki ga obravnavamo, sestavimo množico $\mathcal{A}$, ki ji bomo rekli *abeceda* tega področja. Elementom abecede bomo rekli kar *znaki*. V abecedo bomo dodali tudi znak za *presledek*, ki ga bomo označevali z znakom $\sqcup$, da bo bolj viden. *Niz* ali *beseda* dolžine m je končno zaporedje $\sigma = a_1a_2\ldots a_m$, kjer je vsak $a_k \in \mathcal{A}$, z izrazom *prava beseda* pa bomo imenovali niz, v katerem ni znaka za presledek. Tudi v naravnem jeziku namreč v besedah med črkami ni znaka za presledek. Niz $\sigma = a_1a_2\ldots a_m$ je torej prava beseda, če je v njej vsak $a_k \in \mathcal{A}\setminus\{\sqcup\}$. Pametno je vpeljati tudi znak za *prazen niz*: $\diamond$. Dolžino niza σ označimo z $|\sigma|$. Za $\sigma = a_1a_2\ldots a_m$ je torej $|\sigma| = m$. Katero koli nenegativno celo število m je zato lahko dolžina nekega niza. Tako je na primer $|\diamond| = 0$, $|\sqcup\sqcup| = 2$, $|\texttt{RAKETA}| = 6$ in $|\texttt{RAK}\sqcup\texttt{ETA}| = 7$.

Pravi besedi

$$\sigma = a_1a_2\ldots a_m\,, \quad \tau = b_1b_2\ldots b_n$$

poravnamo v enako dolga niza σ' in τ' z dodajanjem znaka $\sqcup$ tako, da znaki ostanejo v istem vrstnem redu, znak $\sqcup$ pa ne sme biti v σ' in τ' hkrati na isti poziciji. Če bi dovolili nasprotno, bi lahko dobili s poravnavo nize poljubne dolžine s poljubno mnogo presledki. Vzemimo, da sta niza po poravnavi dolžine ℓ:

$$\sigma' = a'_1a'_2\ldots a'_\ell\,, \quad \tau' = b'_1b'_2\ldots b'_\ell\,.$$

Pri tem je a'_k eden od a_i ali $\sqcup$, b'_k eden od b_i ali $\sqcup$, za noben k pa ni $a'_k = b'_k = \sqcup$, števila m, n, ℓ pa očitno povezuje relacija $\max(m,n) \le \ell \le m+n$. Za prikaz poravnave je najbolje, da niza pišemo drugega nad drugim.

Navedimo nekaj poravnav različnih dolžin besed `OBZORNIK` in `MATEMATIKA`:

OB␣␣ZORNIK	OB␣␣ZORNIK␣	OB␣Z␣OR␣N␣IK␣	␣O␣BZO␣RNIK␣
MATEMATIKA	␣MATEMATIKA	␣␣MATEMA␣TIKA	MAT␣EMAT␣IKA

Poravnavo pravih besed σ in τ imamo lahko tudi za pretvorbo besede σ v besedo τ ali obratno s tremi transformacijami nad njunimi znaki, in sicer z brisanjem, vrivanjem in zamenjavo (vključujoč ohranjanje).

Ena od poravnav besed `AKSIOM` in `VZROK` dolžine 8 je

AKSI␣OM␣
␣V␣ZRO␣K

V tem primeru smo po vrsti uporabili naslednjih 8 transformacij:

1. brisanje znaka `A`;
2. zamenjavo znaka `K` z znakom `V`;
3. brisanje znaka `S`;
4. zamenjavo znaka `I` z znakom `Z`;
5. vrivanje znaka `R`;
6. ohranjanje znaka `O`;
7. brisanje znaka `M`;
8. vrivanje znaka `K`.

Po poravnavi dobljena niza spet stisnemo z opustitvijo presledkov:

$$\texttt{AKSI␣OM␣} \mapsto \texttt{AKSIOM}, \ \texttt{␣V␣ZRO␣K} \mapsto \texttt{VZROK}.$$

V tem smislu smo prek poravnave besed z brisanjem, vrivanjem in zamenjavo znakov pretvorili eno besedo v drugo. Podobno lahko opišemo tudi obratno transformacijo, to je pretvorbo besede `VZROK` v besedo `AKSIOM`.

Poravnavo besed lahko vpeljemo tudi drugače. Iz abecede $\mathcal{A}$ konstruiramo abecedo $\mathcal{B} = \mathcal{A} \times \mathcal{A}$. Znake nove abecede $\mathcal{B}$ označimo v obliki stolpca:

$$\begin{bmatrix} a \\ b \end{bmatrix} \in \mathcal{B} \iff a \in \mathcal{A} \quad \text{in} \quad b \in \mathcal{A}\,.$$

Posebej označimo znak za presledek ␣ v $\mathcal{B}$:

$$␣ \in \mathcal{B} \iff ␣ = \begin{bmatrix} ␣ \\ ␣ \end{bmatrix}.$$

Prave besede v $\mathcal{B}$ so po našem dogovoru končni nizi elementov iz $\mathcal{B} \setminus \{␣\}$.

Iz dvojice pravih besed $\sigma = a_1 a_2 \ldots a_m$ in $\tau = b_1 b_2 \ldots b_n$ sestavimo novo besedo dolžine ℓ z znaki abecede $\mathcal{B}$ takole:

$$\begin{bmatrix} a'_1 \\ b'_1 \end{bmatrix} \begin{bmatrix} a'_2 \\ b'_2 \end{bmatrix} \ldots \begin{bmatrix} a'_\ell \\ b'_\ell \end{bmatrix}. \tag{1}$$

Prve koordinate stolpcev so znaki besede σ ali pa ␣. Podobno so druge koordinate znaki besede τ ali ␣. Pri tem je vrstni red znakov brez ␣ v novih nizih enak kot v prvotnih nizih. Nikjer v nizu pa ne sme biti ␣ nad ␣. Če spregledamo oklepaje, dobimo ravno poravnavo besed σ in τ. Torej pri tem nastopajo stolpci treh oblik:

$$\begin{bmatrix} a'_k \\ b'_k \end{bmatrix} = \begin{bmatrix} ␣ \\ b_i \end{bmatrix}, \quad \begin{bmatrix} a'_k \\ b'_k \end{bmatrix} = \begin{bmatrix} a_j \\ ␣ \end{bmatrix}, \quad \begin{bmatrix} a'_k \\ b'_k \end{bmatrix} = \begin{bmatrix} a_i \\ b_j \end{bmatrix}.$$

V poravnavi dolžine ℓ naj bo v število stolpcev prve vrste, h druge in d tretje vrste. Veljajo torej relacije: $v+h+d=\ell$, $m+v=\ell$, $n+h=\ell$. Iz njih izrazimo:

$$d=m+n-\ell\,,\quad h=\ell-n\,,\quad v=\ell-m. \tag{2}$$

Število vseh poravnav dolžine ℓ besed dolžine m in n označimo s $Q(m,n,\ell)$. To število pa lahko zapišemo s *trinomskim koeficientom*. V besedi (1) je namreč nanizanih ℓ stolpcev, od katerih je v prve, h druge in d tretje vrste. Takih pa je toliko, na kolikor načinov lahko permutiramo ℓ reči, od katerih je v prve, h druge in d tretje vrste, zato je

$$Q(m,n,\ell)=\frac{\ell!}{h!v!d!}=\binom{\ell}{h,v,d},\quad h+v+d=\ell.$$

Z upoštevanjem izrazov (2) dobimo:

$$Q(m,n,\ell)=\binom{\ell}{n}\binom{n}{m+n-\ell}=\binom{\ell}{m}\binom{m}{m+n-\ell}. \tag{3}$$

Vseh možnih poravnav pa je

$$D(m,n)=\sum_{\ell=\max(m,n)}^{m+n}Q(m,n,\ell). \tag{4}$$

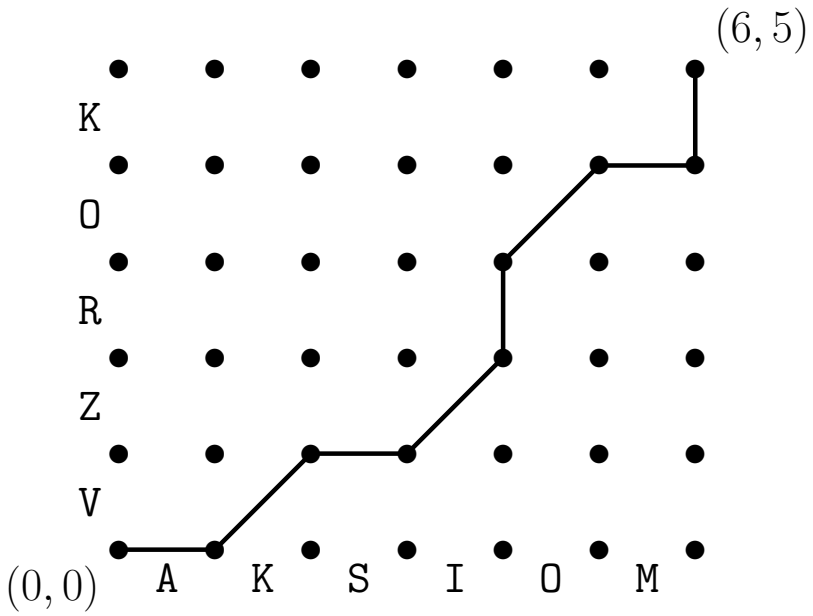
V nadaljevanju bomo z $\mathbb{N}$ označevali množico vseh celih nenegativnih števil, torej $\mathbb{N}=\{0,1,2,3,\ldots\}$. Števila $D(m,n)$ ($(m,n)\in\mathbb{N}\times\mathbb{N}$) imenujemo *Delannoyjeva števila*, ker se je z njimi prvi ukvarjal francoski matematik in častnik *Henri-Auguste Delannoy* (1833–1915) v povezavi s številom šahovskih iger. Več o tem manj znanem matematiku lahko preberemo v [1].

Delannoyjeva števila

Za nazorno ponazoritev poravnav pravih besed dolžin m in n si lahko pomagamo z množico točk

$$\mathcal{M}(m,n)=\{(x,y):x\in\{0,1,2,\ldots,m\},y\in\{0,1,2,\ldots,n\}\}\subset\mathbb{N}\times\mathbb{N}.$$

Vsaki poravnavi pravih besed pa lahko povratno enolično priredimo neko pot, ki povezuje nekatere točke v $\mathcal{M}(m,n)$. Pri tem povežemo točki $(0,0)$ in



Slika 1. Delannoyjeva pot k poravnavi besed `AKSIOM` in `VZROK` iz primera.

(m, n) z neko usmerjeno potjo, pri čemer so dovoljeni samo koraki v smeri vektorjev $(1, 0), (0, 1), (1, 1)$ (horizontalni, vertikalni, diagonalni). Vsaka taka pot je *Delannoyjeva pot*.

Do Delannoyjeve poti, ki ustreza poravnavi pravih besed, pridemo tako, da dani poravnavi priredimo korake takole:

$$\begin{bmatrix} a_j \\ \sqcup \end{bmatrix} \longleftrightarrow (1, 0), \quad \begin{bmatrix} \sqcup \\ b_j \end{bmatrix} \longleftrightarrow (0, 1), \quad \begin{bmatrix} a_i \\ b_j \end{bmatrix} \longleftrightarrow (1, 1).$$

Ustrezajo brisanju, vrivanju in zamenjavi znakov pri pretvorbi ene besede v drugo. Pri tem gremo v poravnavi dolžine ℓ od leve proti desni, ustrezno Delannoyjevo pot pa pričnemo v $(0, 0) \in \mathcal{M}(m, n)$, sledimo poravnavam in po ℓ korakih prispemo v $(m, n) \in \mathcal{M}(m, n)$. Poravnavi besed `AKSIOM` in `VZROK`, ki smo jo navedli zgoraj, ustreza Delannoyjeva pot na sliki 1.

Zato je $Q(m, n, \ell)$ ravno število Delannoyjevih poti z ℓ koraki od $(0, 0)$ do (m, n). Taka pot ima $h = \ell - n$ horizontalnih, $v = \ell - m$ vertikalnih in $d = m + n - \ell$ diagonalnih korakov. Število najdaljših Delannoyjevih poti, ki imajo $m + n$ korakov, je

$$Q(m, n, m + n) = \binom{m + n}{m} = \binom{m + n}{n},$$

število najkrajših Delannoyjevih poti z $\max(m,n)$ koraki pa je

$$Q(m,n,\max(m,n)) = \binom{\max(m,n)}{\min(m,n)}.$$

Število vseh Delannoyjevih poti od $(0,0)$ do (m,n) je ravno Delannoyjevo število $D(m,n)$.

Števila $Q(m,n,\ell)$ so simetrična za vsak $\ell \in \mathbb{N}$: $Q(m,n,\ell) = Q(n,m,\ell)$. Prav tako so tudi Delannoyjeva števila simetrična: $D(m,n) = D(n,m)$. Oboja števila pa imajo rekurzijo. Za števila $Q(m,n,\ell)$ pridemo do nje z naslednjim razmislekom. Denimo, da smo v $\ell - 1$ koraku prispeli iz točke $(0,0)$ v eno izmed točk $(m-1,n), (m,n-1), (m-1,n-1)$. Takih možnosti je $Q(m-1,n,\ell-1) + Q(m,n-1,\ell-1) + Q(m-1,n-1,\ell-1)$. Iz katere koli od omenjenih točk pa je potreben samo še en korak, da prispemo v točko (m,n). Zato za števila $Q(m,n,\ell)$ za vse $\ell \geq 1$, $m \geq 1$ in $n \geq 1$ velja rekurzija

$$Q(m,n,\ell) = Q(m-1,n,\ell-1) + Q(m,n-1,\ell-1) + Q(m-1,n-1,\ell-1)$$

pri robnih pogojih $Q(0,n,\ell) = \delta_{n,\ell}$ in $Q(m,0,\ell) = \delta_{m,\ell}$, kjer je δ znani Kroneckerjev simbol. Posebej namreč za $m > 0$ pomeni $Q(m,0,\ell)$ število poravnav niza dolžine m in praznega niza $\diamond$. Poravnavo, in to dolžine m, v tem primeru dobimo samo na en način, z brisanjem vseh znakov niza σ. Zato je $Q(m,0,m) = 1$. Analogno najdemo: $Q(0,n,n) = 1$. Prazen niz pa poravnamo s praznim nizom tudi le na en način, zato je $Q(0,0,0) = 1$.

Vpeljimo za $\ell \in \mathbb{N}$ množico parov

$$T(\ell) = \{(i,j) \in \mathbb{N} \times \mathbb{N}, 0 \leq i \leq \ell, 0 \leq j \leq \ell, i + j \geq \ell\}.$$

Na tej množici so števila $Q(m,n,\ell)$ pozitivna, na $\mathbb{N} \times \mathbb{N} \setminus T(\ell)$ pa so enaka 0. Zgled je tabela 3. V funkcionalni analizi bi rekli, da je $T(\ell)$ *nosilec* funkcije $(m,n) \mapsto Q(m,n,\ell)$. Pri danem ℓ velja zanimiva enakost

$$\sum_{(m,n)\in T(\ell)} Q(m,n,\ell) = 3^\ell,$$

ki jo dokažemo z metodo matematične indukcije glede na parameter ℓ in s pravkar dokazano rekurzijo, ki jo imajo števila $Q(m,n,\ell)$.

Za Delannoyjeva števila $D(m,n)$ pa velja za vse $m \geq 1$ in $n \geq 1$ rekurzija

$$D(m,n) = D(m-1,n) + D(m,n-1) + D(m-1,n-1) \qquad (5)$$

$n\uparrow$										
8	1	17	145	833	3649	13073	40081	108545	265729	
7	1	15	113	575	2241	7183	19825	48639	108545	
6	1	13	85	377	1289	3653	8989	19825	40081	
5	1	11	61	231	681	1683	3653	7183	13073	
4	1	9	41	129	321	681	1289	2241	3649	
3	1	7	25	63	129	231	377	575	833	
2	1	5	13	25	41	61	85	113	145	
1	1	3	5	7	9	11	13	15	17	
0	1	1	1	1	1	1	1	1	1	
	0	1	2	3	4	5	6	7	8	$m\rightarrow$

Tabela 1. Delannoyjeva števila $D(m,n)$.

pri robnih pogojih $D(m,0)=D(0,n)=1$. Dokažemo jo bodisi iz rekurzije za števila $Q(m,n,\ell)$ in z vsoto (4) ali pa neposredno z naslednjim premislekom.

Denimo, da smo prispeli po Delannoyjevi poti iz točke $(0,0)$ v eno izmed točk $(m-1,n),(m,n-1),(m-1,n-1)$. Takih možnosti je seveda natančno $D(m-1,n)+D(m,n-1)+D(m-1,n-1)$. Iz katere koli od omenjenih točk pa je potreben samo še en dovoljen korak, da prispemo v točko (m,n). Zato za števila $D(m,n)$ velja zgornja rekurzija, s pomočjo katere lahko hitro izračunamo Delannoyjeva števila za majhne m in n (tabela 1).

Obstaja več zapisov števil $D(m,n)$. S formulo (3) na primer dobimo

$$D(m,n)=\sum_{\ell=\max(m,n)}^{m+n}\binom{\ell}{m}\binom{m}{m+n-\ell}=\sum_{\ell=\max(m,n)}^{m+n}\binom{\ell}{n}\binom{n}{m+n-\ell},$$

pa tudi

$$D(m,n)=\sum_{k=0}^{\min(m,n)}\binom{m}{k}\binom{n}{k}2^k\,. \tag{6}$$

Slednjo lahko izpeljemo popolnoma kombinatorično. Vsaki Delannoyjevi poti od $(0,0)$ do (m,n) namreč lahko povratno enolično priredimo neki niz simbolov H,V,D, pri čemer le-ti po vrsti označujejo horizontalni, vertikalni in diagonalni korak. Vzemimo v takem nizu na piko podniza $K=HV$, nekakšen levi ovinek ali koleno, in podniz D. Delannoyjevih poti, na katerih

se K ali D pojavita natanko k-krat, je ravno

$$\binom{m}{k}\binom{n}{k}2^k.$$

Prvi faktor pomeni število načinov, na katere lahko k-krat nastopi eden od omenjenih podnizov po horizontali dolžine m, drugi faktor število načinov, na katere lahko k-krat nastopi eden od omenjenih podnizov po vertikali dolžine n, tretji faktor pa pove, da k-krat izbiramo med dvema podnizoma. Število vseh Delannoyjevih poti potem dobimo kot rezultat seštevanja po vseh možnih številih k.

Pri malo večjih m in n so števila $D(m,n)$ zelo velika, na primer:

$$D(10,10) = 8\,097\,453\,, \quad D(16,24) = 85\,275\,509\,086\,721.$$

Zato je dobrodošla kakšna ocena. A. Raichev in M. C. Wilson sta v [3] izpeljala asimptotično formulo

$$D(Nm,Nn) \sim \left(\frac{n}{L-m}\right)^{Nn}\left(\frac{m}{L-n}\right)^{Nm}\sqrt{\frac{mn}{2\pi NL(m+n-L)^2}}$$

za $N \to \infty$, kjer je $L = \sqrt{m^2+n^2}$.

Preden končamo razdelek, si še oglejmo, kdaj so trinomski koeficienti največji, in lep primer, v katerem se pojavijo Delannoyjeva števila, čeprav ni neposredno povezan z Delannoyjevimi potmi.

Vemo, da binomski koeficienti z izbranim zgornjim indeksom n in spodnjim indeksom k ($n,k \in \mathbb{N}$) pri lihih n dosežejo dvakrat svojo največjo vrednost, pri sodih pa enkrat. Trinomski koeficienti pa pri izbranem zgornjem indeksu dosežejo svojo največjo vrednost enkrat ali trikrat. Za največjo vrednost Q_ℓ trinomskega koeficienta

$$\binom{\ell}{i,j,k} = \frac{\ell!}{i!j!k!} \quad (i,j,k \geq 0, i+j+k=\ell)$$

namreč dobimo izraz

$$Q_\ell = \binom{\ell}{\lfloor\frac{\ell}{3}\rfloor, \lfloor\frac{\ell+1}{3}\rfloor, \lfloor\frac{\ell+2}{3}\rfloor}.$$

Razlikovati je treba tri možnosti glede na ostanek pri deljenju števila ℓ s 3. Označimo z m_ℓ in n_ℓ tisti števili, pri katerih je $Q_\ell = Q(m_\ell, n_\ell, \ell)$. V tabeli 2 so zbrane vse možnosti.

ℓ	m_ℓ	n_ℓ	Q_ℓ
3λ	2λ	2λ	$\frac{\ell!}{\lambda!^3}$
$3\lambda+1$	2λ $2\lambda+1$ $2\lambda+1$	$2\lambda+1$ 2λ $2\lambda+1$	$\frac{\ell!}{\lambda!^2(\lambda+1)!}$
$3\lambda+2$	$2\lambda+1$ $2\lambda+2$ $2\lambda+1$	$2\lambda+2$ $2\lambda+1$ $2\lambda+1$	$\frac{\ell!}{\lambda!(\lambda+1)!^2}$

Tabela 2. Števila Q_ℓ so odvisna od oblike indeksa ℓ.

Da ne bi imeli napačnega vtisa, da smo Delannoyjeva števila vpeljali zgolj zaradi štetja poravnav nizov in Delannoyjevih poti, povejmo, da se pojavljajo še marsikje, o čemer pa lahko več izvemo v izčrpni obravnavi [5]. Omenimo samo primer. Za vsako pozitivno celo število n in vsako nenegativno celo število m je v množici

$$\mathcal{O}_n(m) = \{(x_1, x_2, \ldots, x_n) \in \mathbb{Z}^n : |x_1| + |x_2| + \ldots + |x_n| \leq m\}$$

natančno $D(m, n)$ točk. Podrobnosti najdemo na primer v [4].

$n\uparrow$										
8	1	8	28	56	70	56	28	8	1	
7		8	56	168	280	280	168	56	8	
6			28	168	420	**560**	420	168	28	
5				56	280	**560**	**560**	280	56	
4					70	280	420	280	70	
3						56	168	168	56	
2							28	56	28	
1								8	8	
0									1	
	0	1	2	3	4	5	6	7	8	$m\rightarrow$

Tabela 3. Števila $Q(m, n, 8)$ na nosilcu $T(8)$. Največja so v okvirčkih.

Vemo, da je $\mathbb{R}^n$ poln metrični prostor za normo $||(x_1, x_2, \ldots, x_n)|| = |x_1| + |x_2| + \ldots + |x_n|$ in da so v tem prostoru zaprte krogle s središčem v točki $(0, 0, \ldots, 0)$ in s polmerom $r \geq 0$ množice

$$\mathcal{K}_n(r) = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n : |x_1| + |x_2| + \ldots + |x_n| \leq r\}.$$

Potemtakem množica $\mathcal{O}_n(m)$ vsebuje natanko vse točke s celoštevilskimi koordinatami krogle $\mathcal{K}_n(m)$ in $D(m,n)$ pove število teh točk.

Samo po sebi je zanimivo, kako poimenovati nekatere geometrijske objekte v prostoru $\mathbb{R}^n$. Za $n=1$ imamo opravka kar z množico $\mathbb{R}$, ki jo točka $x_1=0$ razdeli na dva dela: na pozitivna in negativna števila. Krogla $\mathcal{K}_1(r)$ pa je kar daljica s krajiščema v točkah $-r$ in r. Za $n=2$ koordinatni osi $x_1=0$ in $x_2=0$ razdelita ravnino $\mathbb{R}^2$ na 4 *kvadrante*, krogla $\mathcal{K}_2(r)$ pa je običajni kvadrat z oglišči $(\pm r,0)$ in $(0,\pm r)$. Za $n=3$ koordinatne ravnine $x_1=0, x_2=0$ in $x_3=0$ razdelijo prostor $\mathbb{R}^3$ na 8 *oktantov*, krogla $\mathcal{K}_3(r)$ pa je oktaeder z oglišči $(\pm r,0,0)$, $(0,\pm r,0)$ in $(0,0,\pm r)$.

V splošnem primeru koordinatne hiperravnine $x_1=0, x_2=0, \ldots, x_n=0$ razdelijo prostor $\mathbb{R}^n$ na 2^n delov, *ortantov*. Krogla $\mathcal{K}_n(r)$ pa je n-razsežno *hipertelo* z oglišči

$$(\pm r,0,0,\ldots,0),\ (0,\pm r,0,\ldots,0),\ \ldots,\ (0,0,0,\ldots,\pm r).$$

Imena za to hipertelo so različna. Po analogiji s primerom $n=3$ mu pravijo n-razsežni *hiperoktaeder*. V uporabi pa sta tudi imeni *križni politop* in *ortopleks*. Za naravno število m ima torej n-razsežni hiperoktaeder natanko $D(m,n)$ celoštevilskih točk.

Levenštejnova razdalja

Videli smo, da vsaki poravnavi pravih besed $\sigma=a_1a_2\ldots a_m$ in $\tau=b_1b_2\ldots b_n$ v niza $\sigma'=a'_1a'_2\ldots a'_\ell$ in $\tau'=b'_1b'_2\ldots b'_\ell$ ustreza neka Delannoyjeva pot med točkami množice $\mathcal{M}(m,n)$ in obratno. Delu poravnave na j-tem mestu, zapisane s stolpcem oblike

$$\begin{bmatrix} a'_j \\ b'_j \end{bmatrix},$$

pa lahko določimo tudi *kazen* ali *ceno* $c(j,\sigma',\tau')$. Ta kazen je pozitivna, vzeli bomo kar 1, če sta komponenti različni, in 0 sicer. To se pravi:

$$c(j,\sigma',\tau')=\begin{cases} 1\,, & \text{če je} \quad a'_j\neq b'_j\,,\\ 0\,, & \text{če je} \quad a'_j=b'_j\,. \end{cases}$$

Isto lahko zapišemo tudi z relacijo $c(j,\sigma',\tau')=\nu(a'_j\neq b'_j)$, pri čemer $\nu(I)$ pomeni Boolovo vrednost (1 ali 0) izjave I. Celotno kazen ali ceno poravnave pa definiramo kot

$$c(\sigma',\tau')=\sum_{j=1}^{\ell} c(j,\sigma',\tau').$$

Želimo pa si, da je $c(\sigma',\tau')$ najmanjša. Najmanjša kazen poravnave besed σ,τ je *Levenštejnova razdalja* $d_L(\sigma,\tau)$ med njima. Poimenovana je po Vladimirju Iosifoviču Levenštejnu, leta 1935 rojenem ruskem znanstveniku, uveljavljenem na področju teorije informacij in kod za popravljanje napak. Za vsako besedo σ je $d_L(\sigma,\sigma)=0$ in $d_L(\sigma,\diamond)=|\sigma|$. Levenštejnova razdalja

med besedama pove najmanjše število vrivanj, brisanj in zamenjav znakov, ki so potrebni za poravnavo teh besed.

Od poravnav

```
AKSI␣OM␣      AKSIOM
␣V␣ZRO␣K      VZR␣OK
.....=..      ....=.
```

ima prva kazen $c = 7$, druga pa najmanjšo, $c = 5$. Znaki pod poravnavo povejo kazen: pika 1 kazensko točko, enačaj nobene. Zato je v našem primeru $d_L(\texttt{AKSIOM}, \texttt{VZROK}) = 5$. Enako kazen imajo lahko različne poravnave istih besed.

Algoritem za iskanje Levenštejnove razdalje med nizoma najdemo v marsikaterem viru s tega področja, v našem jeziku je lepo opisan v [6]. Zato bomo zgolj navedli ta algoritem za iskanje razdalje $d_L(\sigma, \tau)$ besed

$$\sigma = a_1 a_2 \dots a_m\,, \quad \tau = b_1 b_2 \dots b_n,$$

ki nam da tudi Delannoyjevo pot ustrezne poravnave. Rešitev je lahko več.

Definiramo števila $L(i, j)$, kjer je $0 \le i \le m$ in $0 \le j \le n$. Najprej za $0 \le i \le m$ in $0 \le j \le n$ postavimo

$$L(i, 0) = i, L(0, j) = j. \tag{7}$$

Nato računamo rekurzivno z relacijo

$$L(i, j) = \min(L(i-1, j) + 1, L(i, j-1) + 1, L(i-1, j-1) + \nu(a_i \neq b_j))$$

in na koncu dobimo $d_L(\sigma, \tau) = L(m, n)$. Z uvedbo polkolobarja $(\mathbb{R} \cup \{\infty\}, \oplus, \odot)$, v katerem sta definirani operaciji $\oplus$ in $\odot$ s formulama

$$x \oplus y = \min(x, y) \quad \text{in} \quad x \odot y = x + y,$$

lahko zapišemo

$$L(i, j) = 1 \odot L(i-1, j) \oplus 1 \odot L(i, j-1) \oplus \nu(a_i \neq b_j) \odot L(i-1, j-1), \tag{8}$$

kar spominja na rekurzijo Delannoyjevih števil v obliki

$$D(i, j) = 1 \cdot D(i-1, j) + 1 \cdot D(i, j-1) + 1 \cdot D(i-1, j-1)\,.$$

S polkolobarjem $(\mathbb{R} \cup \{\infty\}, \oplus, \odot)$ se ukvarja *tropska matematika*, ki je dobila ime po vremenskih lastnostih Brazilije, kjer jo je okrog leta 1990 utemeljila skupina francoskih in brazilskih matematikov kot samostojno teorijo.

R	6	5	4	4	3	2	**3**
E	5	4	3	3	2	2	3
T	4	3	3	2	1	2	3
S	3	2	2	1	2	3	4
I	2	1	1	2	3	4	5
S	1	0	1	2	3	4	5
	0	1	2	3	4	5	6
		S	E	S	T	R	A

Tabela 4. Števila $L(i, j)$.

Pogosto je s tem v zvezi omenjen matematik in informatik madžarskega rodu Imre Simon (1943–2009). V tropsko matematiko spadata na primer tropska geometrija in tropska algebra. Tropska matematika je pripravna tudi za obravnavo nekaterih problemov kombinatorike in dinamičnega programiranja, kamor uvrščamo tudi prej opisano iskanje Levenštejnove razdalje med nizoma.

Oglejmo si primer etimološko bližnjih besed `SESTRA` in `SISTER`. Z uporabo robnih pogojev (7) in rekurzije (8) izpolnimo tabelo 4 števil $L(i, j)$.

Torej je $d_L(\texttt{SESTRA}, \texttt{SISTER}) = L(6, 6) = 3$. Od $L(6, 6)$ naredimo obratno Delannoyjevo pot do $L(0, 0)$. Splošno: v levo, če je $L(i-1, j) < L(i, j)$; navzdol, če je $L(i, j-1) < L(i, j)$; diagonalno, če je $L(i-1, j-1) \leq L(i, j)$. V našem primeru imamo dve rešitvi:

R	6	5	4	4	3	2	**3**	R	6	5	4	4	3	**2**	**3**
E	5	4	3	3	2	**2**	3	E	5	4	3	3	**2**	2	3
T	4	3	3	2	**1**	2	3	T	4	3	3	2	**1**	2	3
S	3	2	2	**1**	2	3	4	S	3	2	2	**1**	2	3	4
I	2	1	**1**	2	3	4	5	I	2	1	**1**	2	3	4	5
S	1	**0**	1	2	3	4	5	S	1	**0**	1	2	3	4	5
	0	1	2	3	4	5	6		**0**	1	2	3	4	5	6
		S	E	S	T	R	A			S	E	S	T	R	A

Tabela 5. Ustrezni Delannoyjevi poti sta naznačeni s števili v krepkem tisku.

Ustrezni poravnavi sta:

```
SESTRA        SEST␣RA
SISTER        SISTER␣
=.==..        =.==.=.
```

Poravnava nizov in genetika

Že skoraj 70 let (od leta 1944, veliko o tem je v [7]) je znano, da so v vseh živih organizmih razen v nekaterih virusih nosilke dednih informacij molekule deoksiribonukleinske kisline (DNK). Molekulo DNK si predstavljamo kot dvojno vijačnico, sestavljeno iz riboznih (sladkornih) in fosfatnih gradnikov. Obe vijačnici pa prečno povezujejo pari dušikovih baz adenin (A), citozin (C), gvanin (G) in timin (T). Vedno sta v komplementarnem paru: adenin in timin ter citozin in gvanin. Molekulo DNK si lahko predstavljamo tudi kot lestev, ki smo jo zvili okoli njene vzdolžne simetrale, klini lestve pa so pari dušikovih baz. Tako zgradbo sta leta 1953 na podlagi svojih in predhodnih raziskav drugih znanstvenikov predvidela molekularni biolog in fizik F. H. Crick (1916–2004) in molekularni biolog J. D. Watson (rojen leta 1928), ki sta leta 1962 skupaj s fizikom in molekularnim biologom M. D. Wilkinsom (1916–2004) prejela Nobelovo nagrado za fiziologijo ali medicino. Slednji je z rentgensko metodo potrdil predvidevanje prvih dveh.

Vzdolž izbrane vijačnice molekule DNK si sledijo dušikove baze v nekem zaporedju, ki ga vedno beremo v dogovorjeni smeri, vzdolž preostale vijačnice pa v obratnem komplementarnem zaporedju. Zaporedje lahko obravnavamo kot niz znakov abecede $\{A, C, G, T, -\}$. Črtico $-$ je smiselno pridružiti abecedi, ker z njo v zaporedju označimo *izbris* ali *delecijo* posamezne baze. V nizu baz je zakodiran dedni zapis organizma. Genetiki s posebnim postopkom pridejo do teh nizov, jih poravnavajo in primerjajo med sabo. Tako ugotavljajo na primer evolucijsko sorodnost med organizmi, ki jim DNK pripada.

Povezavo med številom poravnav nizov in Delannoyjevimi potmi ter uporabo v genetiki obravnava veliko del, na primer [2].

LITERATURA

[1] C. Banderier, S. Schwer, *Why Delannoy numbers?*, Journal of Statistical Planning and Inference **135** (11/2005), str. 40–54.

[2] L. Pachter, B. Sturmfels, *The mathematics of phylogenomics*, SIAM Review **49** (2007), št. 1, str. 3–31.

[3] A. Raichev, M. C. Wilson, *A new method for computing asymptotics of diagonal coefficients of multivariate generating functions*, arXiv:math/0702595v1, 9 str.

[4] M. Razpet, *Nekaj primerov kombinatoričnega preštevanja*, Matematika v šoli **13** (2007), št. 1–2, str. 78–96.

[5] R. A. Sulanke, *Objects counted by the Delannoy numbers*, Journal of Integer Sequences **6** (2003), članek 03.1.5, 19 str.

[6] A. Taranenko, *Razdalja med nizi*, Presek **35** (2007/08), št. 1, str. 25–27.

[7] J. D. Watson, A. J. Berry, *DNK – skrivnost življenja*, Modrijan, Ljubljana, 2007.