Tipogenetika

A szövegbányászat egyik érdekes és igen speciális határterülete a tipogenetika. Az elnevezés a tipográfiai genetika terminológia rövidített változata. A tipogenetika a biológiai genetika speciális kiterjesztése, annak karaktersztring alapú tanulmányozása. Elsősorban genetikai, illetve polimertechnológiai területen lehetnek alkalmazásai, de sztringmanipulációs jellege miatt érintőlegesen a szövegbányászat területéhez is kapcsolódik. A téma — amely közel áll a mesterséges élettel kapcsolatos kutatásokhoz és a sejtautomatákhoz — az első publikáció óta (Hofstadter, 2000) egyre nagyobb érdeklődést és kutatási aktivitást vált ki világszerte. A tipogenetika iránti érdeklődést a biológiai genetika eredményei nagy mértékben elősegítették.

A tipogenetikai modell áttekintése

A tipogenetika néhány karakterláncokra vonatkozó szabály által leírt mesterséges rendszer, amely szöveges sztringeken keresztül alkalmazza a genetika eredményeit és módszereit. A szabályokon kívül a rendszerben sztring átalakító műveleteket, ún. mesterséges enzimeket is értelmezünk. A modell formálisan tehát három komponensével írható le: sztringek, enzimek és szabályok (ez utóbbiak tényleges alkalmazásai a műveletek).

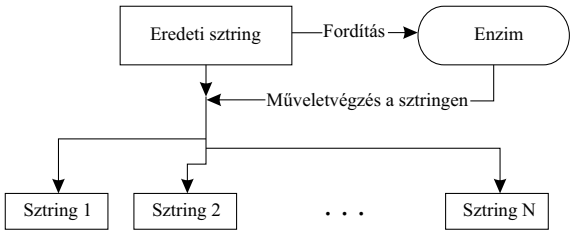
Nézzük részletesebben a tipogenetikai modell ezen alkotóelemeit:

- Sztring: A modell alaphalmazát karakterláncok (sztringek) alkotják. A tipogenetikai sztringek alapesetben a genetikából vett négy ismert karakter (A – adenin, C – citozin, T – timin, G – guanin) által formált karakterláncok.[1]
- Enzim: A tipogenetikai enzimek sztringmanipuláló műveletek sorozataiként értelmezhetőek. Egy sztring önmaga is lehet egy enzim, az ún. fordítás folyamata alakíthat át egy sztringet egy enzimmé. Ezáltal egy sztring tartalmazhatja a saját maga előállításához elvezető műveletek sorozatát is.
- Szabály: A tipogenetikai szabályok a sztringeken értelmezett manipulációk definíciói. A szabályok alkalmazásával valósulnak meg a tipogenetika műveletei, amelyek újabb sztringeket eredményeznek.

A tipogenetikai rendszer alapfolyamata az 1. ábrán látható.

A tipogenetika, ill. a sztringmanipuláció egyik kézenfekvő, érdekes kérdése, hogy előállhat-e olyan helyzet, amikor a leszármazott sztringek között szerepel

[1] Az angol terminológia a strand (szál) szót használja a sztring helyett.



1. ábra. A tipogenetikai rendszer alapfolyamata

az eredeti sztring is, azaz a modell tartalmazza-e az önreprodukciós tulajdonságot. Mint látni fogjuk, a válasz igen: bizonyos tipogenetikai sztringek képesek az önreprodukcióra. Amennyiben a rendszerben lévő sztringekre rekurzív módon alkalmazzuk a szabályrendszert, akkor egy fraktálhoz hasonló eredményhez jutunk, amelyben fellelhető az önhasonlóság, az önhivatkozás és az önreprodukció. Az önreprodukció definíciója szerint egy populáció azon egyedei rendelkeznek önreprodukciós képességgel, amelyek egy szaporodási ciklust követően képesek elérni, hogy a következő generációban is változatlanul jelen legyenek.

A tipogenetikai sztringek azon kívül, hogy a modell alaphalmazának tekinthetőek, az enzimek formájában magukban kódolják a saját magukon elvégzendő műveleteket is, amelyek megvalósítják a szaporodási ciklusokat, és újabb generációkhoz vezetnek. Az újabb generációkhoz vezető műveleteket elvégző enzimek tehát a sztringekbe kódolva implicit módon találhatóak meg a rendszerben. Egy sztring több enzimet is kódolhat önmagában — hasonlóan ahhoz, ahogy a DNS-szál kódolja azokat az enzimeket, amelyek a szaporodásnál elvégzik a DNS-szálon a műveleteket. Az ún. fordítás során egy adott enzim elvégzi az általa kódolt műveletet egy megfelelő sztringen. Egy enzim több sztringre is hathat egyszerre, illetve az enzimműveletek során több sztring is keletkezhet. A fordítás során (amikor a sztringből enzim keletkezik) a sztring megmarad eredeti formájában, ilymódon képes tárolni egy sztring a saját maga átalakítására szolgáló enzime(ke)t.

Az eredeti tipogenetikai rendszer formális felépítése

Az alábbiakban megadott definíciók az elsőként publikált tipogenetikai rendszer modelljét írják le (Hofstadter, 2000). Ettől eltérő tipogenetikai rendszerek is elképzelhetőek más szabályrendszerrel és más karakterkészlettel. Ilyen irányú kutatások találhatóak meg Morris és Varetto írásaiban (Varetto, 1993).

Az alaphalmazt az alábbi négy karakter, $\{A,C,G,T\}$, és a szóköz alkotja, az ebből alkotott karakterláncok a sztringek (pl. *GATTACA_AACCTT*). A tipogenetikai terminológia a karaktereket a genetikai párhuzam miatt *bázis*nak nevezi, az általa elfoglalt pozíció pedig az egység. Például az *ACGGTTA* sztringben a *C* bázis a második egységben található. A bázisok két osztályba sorolhatóak: purinok, illetve pirimidinek. Minden bázisnak van komplementere a másik osztályból, azaz a bázisok bázispárokat alkotnak. A bázispárokat és osztályaikat az 1. táblázat tartalmazza.

1. táblázat. Bázispárok

Purin		Pirimidin
A	$\longleftrightarrow$	T
G	$\longleftrightarrow$	C

A fordítás az a riboszómák által végzett folyamat, amelynek során egy szrtringből enzim vagy enzimek keletkeznek. A fordítás a feldolgozott sztringet érintetlenül hagyja. A folyamat egyirányú: csak sztringekből keletkezhet enzim. A fordítás a sztringben egymás után található bázisokat páronként tekinti. Amennyiben a sztring páratlan bázist tartalmaz és a végén egy bázis marad, akkor azt a fordítás során figyelmen kívül hagyjuk. A fordítási folyamat aminosavakat generál, minden tekintett bázispár egy aminosavnak felel meg. Az aminosavak olyan műveletek, amelyek egy sztringen képesek valamiféle változtatást végrehajtani. Egy enzim tehát tulajdonképpen aminosavak sorozata. Az aminosavak definícióját a 2. ábra tartalmazza (Hofstadter, 2000).

		Második bázis			
		A	C	G	T
Első bázis	A		cut$_S$	del$_S$	swi$_R$
	C	mvr$_S$	mvl$_S$	cop$_R$	off$_L$
	G	ina$_S$	inc$_R$	ing$_R$	int$_L$
	T	rpy$_R$	rpu$_L$	lpy$_L$	lpu$_L$

		Második bázis			
		A	C	G	T
Első bázis	A		cut$_S$	del$_S$	swi$_R$
	C	mvr$_S$	mvl$_S$	cop$_R$	off$_L$
	G	inc$_S$	ing$_R$	int$_R$	ina$_L$
	T	rpy$_R$	rpu$_L$	lpy$_L$	lpu$_L$

2. ábra. Aminosavak. A bal oldali ábrán az eredeti Hofstadter által javasolt aminosavrendszer látható, a jobb oldalon ennek Varetto által módosított változata (Hofstadter, 2000)

Az *AA* bázispárnak speciális szerepe van: ez jelzi a sztringben az enzimek közötti határt, tehát gyakorlatilag a szóközt kódolja. Ennek segítségével lehet egy sztringbe több enzimet is kódolni.

PÉLDA. Tekintsük az *CGCTAATAAGT* sztringet. A fordítást során ebből két enzim keletkezik: a *CGCT* és a *TAAG* szálak által kódolt cop-off és rpy-del enzimek. A sztring végén lévő *T* nem kerül feldolgozásra, mert nincs párja. Vegyük észre, hogy a második *AA* sztringrész nem szóközt kódol, hiszen nem tartoznak egy bázispárba: az első *A* az előző bázispár, *TA* második bázisa, a másik *A* pedig az *AG* bázispár első bázisa. Megjegyezzük, hogy két egymás utáni *AA* sztring esetén nem keletkezik enzim, hanem továbblép az algoritmus.

Az enzimek úgy végeznek műveleteket a sztringeken, hogy azokhoz kapcsolódnak. A művelet kimenete attól függően változhat, hogy hol csatlakozik az enzim a sztringre. Az enzim kötődési preferenciával rendelkezik, amely meghatározza, hogy a sztring mely részeihez csatlakozhat, mielőtt a sztringmanipuláló műveletét megkezdené. Ahogyan a biológiai genetikában beszélhetünk a proteinek másodlagos struktúrájáról, úgy a tipogenetikai rendszerben is értelmezve van az enzimek másodlagos struktúrája. Ez utóbbit az aminosavak csavarodási iránya befolyásolja. A 2. ábrán az *s*, *r* és *l* alsó indexekek rendre arra utalnak, hogy az aminosavnak nincs csavarodása, a csavarodás jobb, illetve bal irányú. Konvenció szerint a vizuális szemléltetésnél az első aminosavat úgy rajzoljuk, hogy a következő aminosav tőle mindig jobbra essen. A kötődési preferenciát az enzim utolsó két aminosava közötti kapcsolat iránya határozza meg. Amennyiben egy enzim csak egy aminosavból áll, tehát a fenti definíció nem alkalmazható rá, akkor megegyezés szerint az *A* bázishoz kapcsolódik. A modell alapfeltételeitől függően a gondolatkísérletet tovább lehet árnyalni azzal, hogy mi történik, ha nincs *A* bázis sem. Ekkor megegyezés szerint általában az enzim nem képes kifejteni semmilyen hatást egyik sztringen sem.

A kötődési preferenciák az utolsó aminosav csavarodási iránya alapján a 2. táblázatban láthatóak.

PÉLDA. A kötődési preferencia meghatározásának szemléltetésére nézzük a 3. ábrán látható példát. Az aminosavak az eredeti aminosavtáblázat alapján kerültek kiválasztásra. Mivel az utolsó link balra mutat, ezért az enzim a *T* bázishoz tud kötődni.

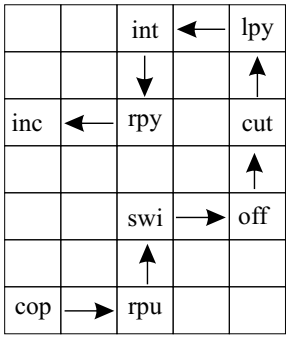
Az eddig ismertetett tipogenetikai rendszer több vonatkozásban is nem determinisztikus. Amennyiben egy sztringre több helyen is rá tud csatlakozni egy enzim, akkor fenti definíciók nem határozzák meg, hogy mi történjék. Több rend-

2. táblázat. Kötődési preferencia

Utolsó link	Kötődési preferencia
$\Rightarrow$	A
$\Uparrow$	C
$\Downarrow$	G
$\Leftarrow$	T

szer képzelhető el attól függően, hogy több lehetséges kapcsolódási bázis jelenléte esetén hogyan definiáljuk az enzimek kötődési szabályát. Amikor egy enzim hozzátapadt egy sztringhez, akkor az enzim minden aminosava elvégzi a műveletét az adott bázison. Amint egy bázissal végzett az enzim, akkor a következőre lép. Az enzim mozgása a sztringen analóg azzal, ahogy a Turing-gép olvasófeje halad végig a szalagon. Konvenció szerint a sztringben azt a bázist, amelyhez éppen hozzátapadt egy enzim, kis betűvel jelöljük. Így a *CAGGCtA* sztring esetében a *T* bázishoz tapadt hozzá éppen egy enzim.

Miközben egy enzim műveletet végez egy sztringen, a megfelelő bázispár hatására életbe lép az ún. másolási üzemmód. Ennek során az éppen olvasás alatt lévő aktuális bázis komplementere generálódik, és tapad hozzá az éppen aktuális bázishoz. Az enzimek az így keletkező komplementer sztringre is átválthatnak, és azon is végezhetnek műveletet. A komplementer sztringet megjelenítéskor fordítottan írjuk az eredeti sztring fölé (ld. pl. a 4. ábrát). Míg az eredeti sztring



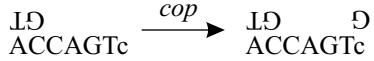
3. ábra. Példa kötődési preferenciára

olvasása balról jobbra történik, addig a komplementer sztring jobbról balra olvasandó. Az ábrán másolási üzemmód esetén kapott eredmény látható egy enzim működése után.

4. ábra. Másolási üzemmód utáni eredmény

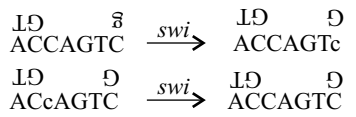
Ez gyakorlatilag három különálló sztringet jelent: *ACCATTGCA*, *GCA*, *GG*. Ahogyan a fordított karaktereket jobbról balra olvassuk, úgy az aminosavak műveleteinek jobb és bal irányai is ennek megfelelően értelmezendőek. Ha egy enzim minden aminosava befejezte a működését az adott sztringen, akkor az enzim leválhat a sztringről. Amennyiben egy szóközre, tehát két bázislánc közé, lép az enzim, ott is megszakad működése az előző bázisszálon, kivéve, ha másolási üzemmódban az rpy, rpu, lpy és lpu aminosavak aktívak. Az alábbiakban ismertetjük az egyes aminosavak működését (ld. még a 2. ábrát).

- cop: Bekapcsolja a másolási üzemmódot, és az éppen aktuális bázis komplementerét a bázis fölé illeszti. Amennyiben a másolási üzemmód aktív, akkor bárhova lép tovább az enzim, a megfelelő bázis komplementerét afölé illeszti. Akárcsak az alapüzemmódban, a szóköz másolási üzemmódnál is leállítja az enzimet. A cop aminosav működését szemlélteti az 5. ábra.



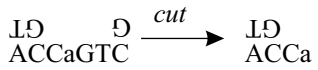
5. ábra. A cop funkció működése.

- off: Kikapcsolja a másolási üzemmódot. Amennyiben a másolási üzemmód nem aktív, így ennek az aminosavnak nincs hatása. Ez a művelet soha nem változtatja meg az éppen feldolgozás alatt lévő bázisszálat.
- swi: Megszünteti a kötődést az eredeti bázisszállal, és a komplementer bázisszálhoz kapcsolja az enzimet. Az enzim tehát átvált az eredeti és a komplementer bázisszál között. Amennyiben nincs komplementer bázisszál, így az enzim befejezi működését. A swi aminosav működését a 6. ábra szemlélteti, ahol az alsó esetben az enzim leáll.
- cut: Az enzim elvágja az éppen aktuális bázistól jobbra mind az eredeti, mind pedig a komplementer bázisszálat. A levágott sztringeket az enzim ezután már nem éri el. Amennyiben az éppen aktuális bázistól jobbra már nincs



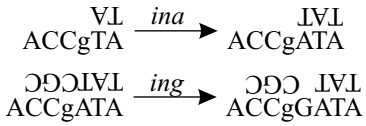
6. ábra. A swi funkció működése

másik bázis, akkor ennek az aminosavnak nincs hatása. A cut aminosav működésére a 7. ábra mutat példát.



7. ábra. A cut funkció működése

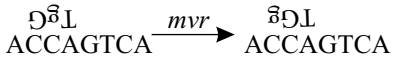
- del: Törli az éppen aktuális bázist, és eggyel jobbra lépteti az enzimet a bázisszálon. Amennyiben a törölt bázistól jobbra már nincs több bázis, akkor az enzim leáll. A törlés nem érinti a komplementer bázisszálat, azonban ha az enzim másolási üzemmódja aktív, akkor a jobbra léptetést követően az új bázis komplementere a bázisszálhoz az új bázis fölé illesztődik. Alapesetben a kitörölt bázis helyére szóköz kerül, de elképzelhető más szabály is, pl. a sztring teljes jobbra lévő része eggyel balra tolódik. Ekkor nem keletkezik szóköz a kitörölt bázis helyén.
- ina, inc, ing, int: Ezek az aminosavak az éppen aktuális bázisszál-pozíció után rendre beszúrják az *A*, *C*, *G* és *T* bázisokat. Amennyiben a másolási üzemmód aktív, akkor a komplementer bázisszálba beszúrásra kerül a beszúrt bázis komplementere. Ha a másolási üzemmód nem aktív, akkor a komplementer sztringbe egy szóköz kerül. A 8. ábra az ina (másolási üzemmód) és az ing (normál üzemmód) aminosavak működését illusztrálja.



8. ábra. Az ina funkció működése

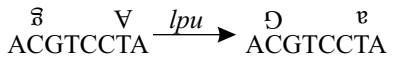
- mvl, mvr: Hatásukra az enzim az éppen feldolgozás alatt álló bázisszálon rendre eggyel balra (mvl), illetve jobbra (mvr) lép. Amennyiben a másolási üzemmód aktív, így a lépés után az újonnan kötődött bázis komplementere a bázisszál fölé az új bázispozíció fölé íródik. Amennyiben az enzim egy szó-

közre vagy a sztring végére lép, az enzim leáll. Az mvr aminosav működésére a 9. ábra mutat példát. Itt egy újabb mvr esetén az enzim leállna.



9. ábra. Az mvr funkció működése

- lpy, lpu: Ezek az aminosavak az éppen aktuális pozíciótól balra megkeresik rendre az első pirimidint, illetve purint, és odakötik az enzimet. Alapértelmezésben nincs megadva, hogy a keresés képes-e a szóközöket átugrani. Az lpu aminosav működését a 10. ábra szemlélteti.

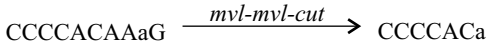


10. ábra. Az lpu funkció működése

- rpy, rpu: Ezek az aminosavak az lpy és lpu aminosavakhoz analóg módon működnek, csak az első pirimidin, ill. purin keresését az aktuális bázis pozíciójához képest jobbra végzik.

Egy szóközökkel elválasztott bázisszálakból álló sztring azokat az enzimeket kódolja, amelyekkel a sztringet fel kell dolgozni. Ekkor a kiinduló sztringből fordítás után létrejönnek azok az enzimek, amelyek utána az eredeti sztringen elvégzik a vonatkozó műveleteket és új sztring-leszármazottakat hoznak ezzel létre. Megegyezés kérdése, hogy a létrejött enzimek milyen sorrendben kerüljenek sorra. Az új sztringek kódolt formában szintén magukban hordozzák azokat az enzimeket, amelyek fordítás után a saját feldolgozásukat szabályozzák. Ily módon a tipogenetikai rendszerben a sztringek evolúciója több generáción keresztül folyhat.

PÉLDA. Tekintsük a *CCCCACAAAG* sztringet, amely az mvl-mvl-cut és a del aminosavakat kódolja. Ekkor az mvl-mvl-cut aminosav hatására, amelynek a kötődési preferenciája *A*, a 11. ábrán látható folyamat történik feltéve, hogy a legutolsó *A*-hoz kötődik az enzim induláskor.



11. ábra. Az mvl funkció működése

Az *AAG* bázisszál a folyamat melléktermékének tekinthető. Kérdés, hogy a del aminosav melyik sztringen kezdje el működését. További döntési lépést jelen-

tene, ha lenne komplementer bázisszál is. Ezen kérdések tisztázása, ill. a megfelelő alternatívák kiválasztása megegyezés kérdése, ettől függően eltérő tipogenetikai rendszerek jönnek létre, különböző eredményekkel.

Tipogenetikai sztringek tulajdonságai

A leszármazott sztringeket úgy kapjuk, hogy az eredeti sztringek általuk kódolt enzimeket végrehajtjuk a kódot tartalmazó sztringre. A leszármazott sztringek alapján értelmezzük a tipogenetikai sztringek különböző tulajdonságait. Például ha a leszármazott sztringek között van olyan, amelyik megegyezik az eredetivel, akkor az eredeti sztring *önreprodukció*s képességű. A sztringek leszármazása egy körmentes irányított gráffal szemléltethető, ahol a csomópontok a bázisszálak, a köztük lévő él pedig a leszármazási viszonyt (gyerek–szülő) fejezi ki, azaz az enzimműveletek folyamatát. A leszármazottjaik tulajdonságai alapján az sztringek az alábbi osztályokba sorolhatók:

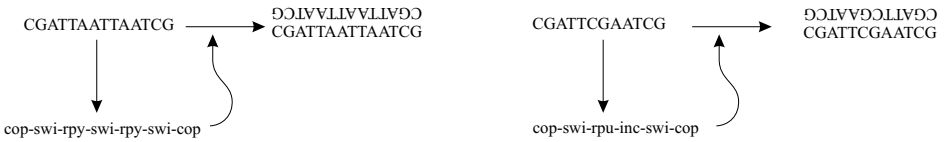
1. *Meddők* osztálya (dud). Ide tartoznak azok a sztringek, amelyek nem képesek leszármazott sztringek előállítására. Ez például akkor fordulhat elő, ha a sztring olyan enzim(ek)et kódol, amely(ek) nem képes(ek) az eredeti sztringhez kötődni. Ilyen például a *CGGC* bázisszál, ami a cop-inc enzimet kódolja. Ennek azonban *A* a kötődési preferenciája, így nem tud a sztringre kötődni.
2. *Önátörökítők* osztálya (self-perpetuators). Ide azok a sztringek tartoznak, amelyek az enzimműveletek rekurzív alkalmazásai során folyamatosan vagy periodikusan jelen vannak a rendszerben a sztringek között, de soha nem fordul elő belőlük másolat, csak mindig egy példány, azaz a sztring periodikusan van jelen adott generációkban. Az osztály speciális alosztályát azon sztringek képezik, amelyek amellett, hogy önmagukat átörökítik a következő generációba, még olyan leszármazott sztringe(ke)t is eredményeznek, amelyek szintén önátörökítőek. Példa lehet az önátörökítő sztringre a *TCCGCAATTT* bázisszál, amely a rpu-cop-mvr-swi-lpu enzimet kódolja. Az enzim létrehoz egy másik sztringet, de az eredeti sztringet sértetlenül hagyja.
3. *Szaporodók* osztálya (self-replicators). Ide azok a sztringek tartoznak, amelyek amellett, hogy önátörökítőek, a későbbi generációk során további másolatokat készítenek magukból biztosítva szaporodásukat a rendszerben.

PÉLDA. Tekintsük a *CGCGCGCGTAATATAACGATCGCGCGTATTAATTAATACGCGCGATCGTTATATTACGCGCGCG* szaporodó sztringet, amely négy enzimet kódol, rendre *C, G, C* és *A* kötődési preferenciákkal:

1. *CGCGCGCGTAATAT*: Az első enzim balról az első *C* bázishoz kötődik, és másolási üzemmódra váltva az első három bázishoz komplementer bázispárokat rendel a komplementer bázisszálon.
2. *CGATCGCGCGTATT*: A második enzim a jobbról az első *G*-hez kötődik, és beilleszt föléje egy komplementer *C*-t másolási üzemmódban, majd átvált a komplementer bázisszálra. Ezt követően az enzim a teljes sztring hosszában másolási üzemmódban minden eredeti bázis fölé beilleszti komplementerét. Ennek eredményeként a komplementer bázisszál az eredeti bázisszállal éppen megegyező lesz.
3. *TT*: Hatástalan.
4. *TACGCGCGATCGTTATATTACGCGCGCG*: Hatástalan.

Az utolsó két enzim változatlanul hagyja a sztringeket. Végül a komplementer és az eredeti sztring kettéválik, ezáltal két teljesen egyforma bázisszál keletkezik. Vegyük észre, hogy az eredeti sztring második fele az első fél komplementere. A biológiában az ilyen tulajdonságú szálakat invertált másolatoknak nevezik.

Egy másik példa szaporodó sztringre a *CGTTTTTTTG* karakterlánc. Ez úgy képes szaporodni, hogy először előállítja saját komplementerét (*CAAAAAAACG*), majd ezt követően az eredeti sztring enzimje szükséges ahhoz, hogy a leszármazott sztringből ennek komplementerét, azaz az eredetivel megegyező sztringet generálja. Döntés kérdése, hogy megengedjük-e azt, hogy egy enzim ne csak arra a sztringre hasson, ami őt kódolta, hanem bármelyik másikra, így a leszármazott sztringekre is. A 12. ábrán további két szaporodó sztringre látható példa (Hofstadter, 2000; Varetto, 1993).



12. ábra. Példa szaporodó sztringre

Forrás:

D. R. Hofstadter. *Gödel, Escher, Bach — Egybefont gondolatok birodalma.* Typotex, 2000.

L. Varetto. Typogenetics: An artificial genetic system. *J. of Theoretical Biology*, 160:185–205, 1993.