Csiszár Villő–Keszei Ernő
ELTE TTK Matematikai Intézet ELTE TTK Kémiai Intézet
villo@cs.elte.hu keszei@chem.elte.hu

Információ a molekulák világában

Bevezetés

A Magyar Kémikusok Lapja 2009. márciusi számában jelent meg „A biokémiai információ hierarchiája” című cikk [1], aminek megértésével problémáink voltak. A cikk az információ, illetve entrópia fogalmát számos értelemben használja, de nem pontosítja, hogy mikor használja hétköznapi, és mikor matematikai értelemben. Néhány kapcsolódó fogalmat (információ összenyomhatósága, rendezettség, bonyolultság) is pontatlanul használ. Fontosnak tartjuk ezért tisztázni a molekuláris információ értelmezéséhez a szükséges információelméleti alapokat, eközben pontosítunk néhány kifejezést is, annak megfelelően, ahogy a jelenlegi magyar szóhasználatban azok előfordulnak. A cikk terjedelmi korlátai miatt ez csak a legszükségesebb ismeretek áttekintésére korlátozódik, lehetőség szerint azonban igyekszünk megkímélni az olvasót bizonyítatlan állításoktól.

Az információ és az entrópia matematikailag precízen definiálható fogalmak, sőt, több különböző mérőszámuk létezik. Említhetjük itt a Shannon-entrópiát, a Hartley-entrópiát, a Kolmogorov-bonyolultságot, vagy a termodinamikai entrópiát. Az [1] közlemény anyagán nem túlmutató áttekintésben mi elsősorban az entrópia különböző meghatározásaival, illetve a források zajmentes átvitelben betöltött szerepével foglalkozunk.

Az [1] cikk a biokémiai információval kapcsolatban három fő irányban folytat vizsgálatokat: 1. miből és hogyan keletkezett a biológiai információ? 2. hogyan számszerűsíthetjük a biológiai információt? 3. a biokémiai információ minősége. Mindhárom kérdés érdekes és fontos. Mi ezek közül a másodikra térünk ki részletesebben, mivel csak ezzel kapcsolatban egyértelmű az információelmélet alkalmazhatósága. Az első kérdésre vonatkozóan csak néhány rövid észrevételt teszünk, míg a harmadik kérdés nem tanulmányozható az információelmélet eszközeivel (az elmélet alap-axiómája, hogy az információnak csak a mennyiségével foglalkozik, a minőségével, értelmével nem).

A Shannon-féle információ és entrópia

A modern információelmélet megteremtőjének Claude E. Shannont tekintik, aki hírközlési rendszerekben vizsgálta az információátvitel elérhető sebességét [2]. A hírközlési rendszerben az *adó* valamely forrásból származó információt szeretne a *vevőnek* továbbítani, egy *csatornán* keresztül. Az adó az információt először a csatornán továbbítható formára hozza, azaz kódolja. A kódolás során az információt tömörítheti is, ezáltal az információ továbbításakor idő és költség takarítható meg. Amennyiben a csatorna zajos, a kódolásnak arra is alkalmasnak kell lennie, hogy a zaj torzító hatásait kiküszöbölje. Shannon felfogásában tehát „az információ üzenet, mely valamilyen jelrendszer segítségével továbbítható”. Amennyiben a csatorna zaj nélküli, azaz a továbbításkor semmiféle hiba nem következhet be, akkor az információátvitel elérhető legnagyobb sebességét az információforrás Shannon-féle entrópiája határozza meg.

Az információforrás tulajdonképpen egy véletlen jelenség vagy kísérlet. A jelenséget csak megfigyeljük, míg a kísérletet mi magunk végezzük el: a biológiában inkább az előbbi fordul elő, így a továbbiakban csak véletlen jelenségekről beszélünk. Az információ onnan származik, hogy megfigyeljük a véletlen jelenség kimenetelét. Ha az adó már ismeri ezt a

kimenetelt, akkor mondhatjuk, hogy információval rendelkezik, melyet szeretne a vevőhöz is eljuttatni. Ha a jelenségnek m lehetséges kimenetele van, és az i-edik kimenetel valószínűsége p_i, akkor a jelenség *Shannon-féle entrópiája*:

$$H = -\sum_{i=1}^{m} p_i \log_2 p_i \text{ bit.} \tag{1}$$

A H mennyiség azt fejezi ki, hogy mennyire bizonytalan a jelenség vagy kísérlet kimenetele. Shannon a logaritmust kettes alapúnak, ezáltal az entrópia mértékegységét bit-nek választotta (a bit a „binary digit", azaz a bináris számjegy rövidítése). A legegyszerűbb eset ugyanis az, amikor csak kétféle kimenetel lehetséges, melyek valószínűségei p és $1 - p$. Ha $p = 0$ vagy $p = 1$, akkor nincs bizonytalanság: az entrópia nulla. Ezzel szemben akkor a legbizonytalanabb a jelenség kimenetele, ha $p = 0{,}5$, ekkor az entrópia 1 bit.

Ha megfigyeljük, hogy az i-edik kimenetel következett be, akkor $-\log_2 p_i$ bit információhoz jutunk. Ezt nevezzük a kimenetel egyedi információjának, ami körülbelül azt fejezi ki, hogy mennyire meglepő ez a kimenetel. Az egyedi információ várható értéke, az átlagos információ tehát:

$$I = -\sum_{i=1}^{m} p_i \log_2 p_i \text{ bit.} \tag{2}$$

Látható, hogy $I = H$, vagyis az entrópia és az átlagos információ megegyezik. Ez világos: gondoljunk arra, hogy az információ azt fejezi ki, hogy a megfigyelés mennyivel csökkenti a bizonytalanságot. A jelenség megfigyelése előtt a bizonytalanság H, a megfigyeléssel azonban teljes bizonyosságot nyerünk, a bizonytalanság nullára csökken.

Ha X jelöli a forrás véletlen kimenetelét, Y pedig egy tetszőleges másik véletlen jelenség kimenetele, akkor megkérdezhetjük, hogy Y mennyi információt tartalmaz X-ről, azaz Y megfigyelése mennyivel csökkenti X bizonytalanságát. Ezt a két jelenség $I(X : Y)$ *kölcsönös információja* fejezi ki. Az előbbiek alapján $I(X : X) = H(X)$, általában azonban $I(X : Y) < H(X)$. Ashby törvénye szerint [3] „Egy determinisztikus fizikai rendszer kimenetének változékonysága nem lehet nagyobb a bemenet változékonyságánál; a kimenet információja nem haladhatja meg a bemenetben jelen lévő információt." Shannon elméletével ezt úgy fogalmazhatjuk meg, hogy ha f egy determinisztikus függvény, akkor $I(X : Y) \geq I(X : f(Y))$, vagyis Y több információt tartalmaz X-ről, mint az $f(Y)$ függvény.

Megjegyzendő még, hogy időnként szokás az (1) képletben természetes alapú logaritmussal számolni; ekkor az entrópia mértékegysége a nat („natural digit") lesz, az átváltás szabálya 1 bit = ln 2 nat = 0,693 nat, illetve 1 nat = 1 / ln 2 bit = 1,443 bit .

Shannont megelőzően Hartley [4] is bevezetett egy mérőszámot az információ mennyiségére. Szerinte egy m lehetséges kimenetelű jelenség megfigyelésekor $\log_2 m$ bit információt nyerünk. A Hartley által definiált információmérték megegyezik a Shannon-félével, amennyiben mindegyik kimenetel egyformán valószínű (azaz $p_i = 1/m$). A következő szakaszban látni fogjuk, hogy az információ továbbításának vizsgálatakor a Shannon-féle entrópia bizonyul hasznosnak. A Hartley-entrópia akkor használható, ha az információforrásról csak azt tudjuk, hogy m lehetséges kimenetele van, de ezek valószínűségeit nem ismerjük. Ekkor az információforrás entrópiája maximálisan $\log_2 m$ bit – éppen a Hartley-entrópia – lehet.

Információforrások blokkonkénti kódolása

Az előző szakaszhoz képest egy lépéssel továbbmenve, vizsgáljuk azt az esetet, amikor az információforrásban a véletlen jelenség nem csak egyszer, hanem sokszor, mondjuk n-szer játszódik le. Ekkor a kimenetelek száma összeszorzódik, azaz m^n kimenetel-sorozat

lehetséges, melynek Hartley-entrópiája $n \log_2 m$ bit. Nevezzük a továbbiakban ezeket a kimenetel-sorozatokat *üzeneteknek*.

Az információelmélet egyik alapkérdése, hogy az üzenetek mennyire tömöríthetők, azaz mennyire „nyomhatók össze" anélkül, hogy információt veszítenénk. Például az n nukleotid hosszúságú DNS-szakaszokat lehet-e rövidebb sorozatokkal kódolni? A válasz természetesen függ attól, hogy milyen ábécét használunk a kódoláshoz. Tegyük fel, hogy a kódoló-ábécé M betűt tartalmaz, és az üzeneteket N hosszú kódszavakkal szeretnénk kódolni (a kódszavakból áll össze a kódszótár, melyről megköveteljük, hogy minden üzenet kódszava különböző legyen). A kódolandó üzenetek száma 4^n (mindegyik nukleotid A, C, G, T lehet), amiből $M^N \geq 4^n$, vagyis $N \geq n / \log_4 M$ kell legyen. Látható, hogy bővebb ábécé felhasználásával rövidebb kód készíthető, ezt azonban nem akarjuk tömörítésnek nevezni. Ezért szokás a kételemű {0, 1} ábécét rögzíteni, azaz bináris kódokra szorítkozni.

Tegyük ismét fel, hogy az üzenetek egy m betűből álló ábécé feletti n hosszúságú sorozatok. Természetes ötlet, hogy az üzeneteket bontsuk fel b hosszúságú blokkokra, és a blokkokat egyenként kódoljuk: ekkor csak m^b darab kódsorozatot kell a kódszótárban nyilvántartani. Jelölje a b hosszúságú blokkok entrópiáját H_b. Megmutatható, hogy a b hosszúságú blokkok legrövidebb bináris (prefix) kódjának átlagos kódszóhossza körülbelül H_b, azaz egy betűre átlagosan H_b / b kódbit jut. Ha az információforrásra létezik a H_b / b mennyiség $\boldsymbol{H}$ határértéke, amint b végtelenhez tart, akkor elég hosszú blokkokat kódolva, az üzenet egy betűjét átlagosan $\boldsymbol{H}$ bittel tudjuk kódolni ($\boldsymbol{H}$ a forrás *betűnkénti entrópiája*). Ha az információforrásban ugyanaz a véletlen jelenség játszódik le sokszor és egymástól függetlenül (ez az emlékezet nélküli stacionárius forrás), akkor $\boldsymbol{H}=H_1$, ha viszont az egymás utáni jelenségek kimenetelei között összefüggések vannak, akkor a forrás betűnkénti entrópiája kisebb, mint egy jelenség entrópiája.

A Kolmogorov-bonyolultság

Az előbbiek szerint, ha egy információforrás betűnkénti entrópiája $\boldsymbol{H}$, akkor a forrásból származó n hosszú sorozatokat $n\boldsymbol{H}$ bittel tudjuk kódolni, azaz leírni. Bizonyos szempontból azok a források tekinthetők bonyolultnak, melyek leírásához sok bitet kell felhasználni, azaz entrópiájuk nagy. Az ilyen források által kibocsátott sorozatokban nincsen szabályszerűség, rendezettség. Ezen az ötleten alapszik a *Kolmogorov-bonyolultság fogalma* (melyet elsőként Solomonoff [5] vezetett be): egy adott sorozat Kolmogorov-bonyolultsága azt fejezi ki, hogy milyen hosszú az a legrövidebb program, mely futási eredményként éppen az adott sorozatot írja ki (a Kolmogorov-bonyolultságot szokás *algoritmikus bonyolultságnak* is nevezni). A pontos definícióhoz természetesen definiálni kellene, hogy milyen programokat engedünk meg, és ezek milyen számítógépen futnak. Shannon-entrópiája egy információforrásnak van, míg Kolmogorov-bonyolultsága egy konkrét sorozatnak, mégis, a két fogalom szoros kapcsolatba hozható, ám erre a kapcsolatra most nem térünk ki. Illusztációként inkább nézzünk két 55 hosszúságú DNS-sorozatot, melyeket egy-egy információforrás állított elő:

1. CAATTTTAGGGTAGCAGACGCACTAGCCGAATATGTTATCTACCTCTCCCCCCCG
2. TGCATGCATGCATGCATGCAT**T**GCATGCATGCATGCATGCATG**G**CATGC**C**ATGCA

Az első sorozat teljesen véletlenszerű, ennek megfelelően nincs rövid leírása (a forrás entrópiája 2 bit). A második sorozat úgy keletkezett, hogy a TGCA mintázatot ismételtük, de minden bázist 10% eséllyel megdupláztunk. A sorozat egy rövid leírása: „12×TGCA, majd duplázd a 21., 42., 47. betűket" (a forrás entrópiája $-(0{,}1 \times \log_2 0{,}1 + 0{,}9 \times \log_2 0{,}9) = 0{,}47$ bit).

A termodinamikai entrópia

A fenomenologikus termodinamika állapotfüggvénye, az S entrópia a statisztikus fizika összefüggései alapján:

$$S = -k\sum_i p_i \ln p_i, \tag{3}$$

ahol p_i a termodinamikai rendszert reprezentáló sokaságok diszkrét állapotainak valószínűségi sűrűségfüggvénye, a k Boltzmann állandó pedig az entrópia skáláját határozza meg, amely Kelvin egységekben mért hőmérséklet és Joule egységekben mért energia esetén $k = 1.3806504\times10^{-23}$ J / K. (A T hőmérséklet megszorozva az S entrópiával energiát ad.)

Mikrokanonikus sokaságban – ami az elszigetelt rendszer reprezentációja, azaz állandó energiájú, állandó térfogatú és állandó összetételű – a feltételeket megvalósító sokaság minden egyes állapota azonos valószínűségű. Ha az állapotok száma Ω, akkor minden egyes állapot valószínűsége $p_i = 1/\Omega$, így az entrópia $S = k \ln \Omega$. Az Ω mennyiséget szokás mikrokanonikus állapotösszegnek vagy partíciós függvénynek is nevezni.

Kanonikus sokaságban – ami a zárt, merev falú, termosztált (azaz állandó hőmérsékletű) rendszer reprezentációja – a feltételeket megvalósító sokaság állapotai nem azonos valószínűségűek, hanem a Boltzmann valószínűség-eloszlással

$$p_i = \frac{1}{Q} e^{-\frac{E_i}{kT}} \tag{4}$$

írhatók le, ahol a Q mennyiség a kanonikus partíciós függvény, ami az exponenciálisok összege az összes lehetséges állapotban:

$$Q = \sum_i e^{-\frac{E_i}{kT}}, \tag{5}$$

az E_i pedig a termosztáttal érintkező rendszer lehetséges energiája. Az entrópia ebben az esetben a kanonikus sűrűségfüggvény logaritmusának várható értékével számítva

$$S = -k\sum_i p_i \ln\frac{1}{Q} e^{-\frac{E_i}{kT}} = k\sum_i p_i \ln Q + k\sum_i p_i \frac{E_i}{kT} = k\ln Q + k\frac{1}{kT}\sum_i p_i E_i = -\frac{F}{T} + \frac{U}{T},$$

mivel az első tag kifejezhető az ismert $F = -kT \ln Q$ összefüggésből, a második tagban pedig az energia várható értéke jelenik meg, amit U-val jelöltünk.

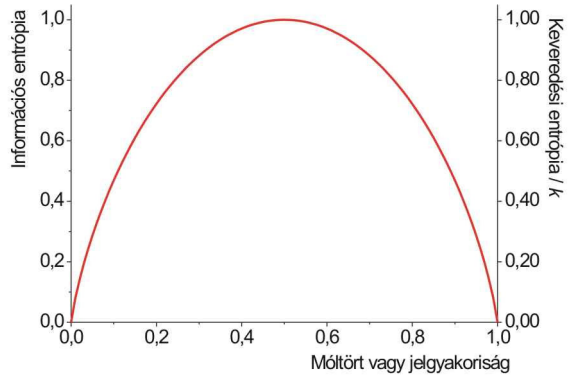
Megállapítható tehát, hogy a skálafaktortól eltekintve – ami csak az entrópia megfelelő egységei miatt szerepel a képletekben – az (1)-beli H Shannon-entrópia pontosan megfelel a (3)-beli S termodinamikai entrópiának. (A neve is innen származik.) A termodinamikai entrópia megfeleltetése az információelméletinek a legtöbb esetben nem egyszerű, de van egy igen szemléletes eset. A keveredési entrópia ideális kétkomponensű elegyben egy részecskére vonatkoztatva az ismert

$$S_{\text{mix}} = -k\left[x_1 \ln x_1 + (1-x_1)\ln(1-x_1)\right]. \tag{6}$$

A kétkomponensű elegyből származó „üzenetekben” az 1-el indexelt molekulák előfordulási valószínűsége éppen $p_1 = x_1$, míg a 2-vel indexelteké $p_2 = (1 - p_1)$. A kételemű eseménytér H Shannon-féle entrópiája így

$$H = -\left[p_1 \log_2 p_1 + (1-p_1)\log_2(1-p_1)\right]. \tag{7}$$

Mivel a 2-es alapú logaritmus és a természetes alapú logaritmus skálái között van egy ln 2 átváltási faktor, ezért a k-val elosztott S_{mix} keveredési entrópia a H entrópiától már csak az ln 2 szorzóban különbözik. Amint az az 1. ábrán látható, a megfelelően skálázott entrópiák azonosak.



1. ábra. Az ideális kétkomponensű elegy (6) összefüggés alapján számítható keveredési entrópiája (jobb oldali skála) és a két jelből álló jelkészlet (7) összefüggés alapján számítható információs entrópiája (bal oldali skála) az x_1 móltört, illetve a p_1 jelgyakoriság függvényében.

Érdekes megjegyezni, hogy a két komponens keveredésekor kinyerhető F_{mix} szabadenergia éppen TS_{mix} , ami azt is jelenti, hogy a részecskék szétválasztásához (kvázisztatikus esetben) szükséges energia is ekkora, ami éppen a T hőmérséklet szorzata a H Shannon-féle entrópiával – megfelelő skálázás esetén. Ezért szokás azt mondani, hogy a részecskék „szétválogatása” csak energia befektetésével lehetséges, amely energia arányos a szétválogatáshoz szükséges információval.

A termodinamika második főtétele szerint egy elszigetelt rendszer entrópiája nem csökkenhet, azaz lokálisan előfordulhat entrópiacsökkenés, de csak úgy, hogy a rendszer más pontjain nő az entrópia. Mint arra az [1] közlemény is rámutat, az élet keletkezése nem mond ellent a termodinamika második főtételének, hiszen a rendezettség lokálisan (a Földön) nőhet, ha ezt az univerzumban kiegyenlíti a rendezetlenség növekedése. (Ehhez még az is szükséges, hogy az entrópiához egy „rendezetlenség” értelmet is társítsunk, ami csak meglehetősen korlátozott érvénnyel tehető meg.)

A biológiai információ

Az élőlények egyedfejlődése, élettani folyamatai, sőt viselkedési mintái is – legalább részben – az őseiktől örökölt információ alapján alakulnak. Ezen információ legnagyobb, bár korántsem teljes része a DNS-ben tárolódik. A szülők továbbítják utódaiknak a létrejöttükhöz és működésükhöz szükséges leírást, azaz információt, a DNS csatornáján és kódján keresztül. A DNS molekula csodálatosan alkalmas arra, hogy egy bonyolult élőlény felépítéséhez és működéséhez szükséges rengeteg információt tárolja, illetve megbízható módon örökítse. Ennek fontos eleme, hogy az élőlények bonyolultságával arányosan változhat a molekula hosszúsága, így az általa kódolható információtartalom is. Az emberi genom kb. 3,2 milliárd ($3{,}2 \times 10^9$) bázispárból áll, ami 23 kromoszómában oszlik meg, így ezek átlagosan 160 millió bázispárt tartalmaznak. Ha ezek egybetűs jeleit sorra beírnánk egy könyv lapjaira, az kb. 100 000 oldalra férne csak ki.

Felmerül a kérdés, hogy valójában mekkora a genom által kódolt információ mennyisége? A DNS-ben a biológiai információt a bázispárok sorrendje hordozza. Ha tehát a DNS-re (és egyéb lineáris szerkezetű polimer molekulákra, pl. RNS-re vagy fehérjékre), mint egy ábécé betűiből felépített jelsorozatokra tekintünk, akkor a bennük rejlő információ mérésére a termodinamikai entrópia semmiképpen sem alkalmas. Ezért félrevezető az [1] közlemény 2.

táblázata, melyben a Hartley-féle és a termodinamikai entrópia keveredik. Így az alábbiakban csak a Hartley-féle entrópia kiszámításával foglalkozunk.

Mivel a DNS-ben az A, C, G, T betűk fordulnak elő, egy betű maximum $\log_2 4 = 2$ bit információt hordozhat. Így egy $4 \cdot 10^6$ hosszúságú modell-DNS maximálisan $8 \cdot 10^6$ bit információt hordozhat. (Ez a DNS-méret egyébként azonos az *E. Coli* baktérium genomjának méretével.) Figyelembe vehetjük azonban, hogy a genetikai kód a lehetséges 64 bázishármasból az aminosavak kódolására csak 61-et használ. Ezek szerint egy bázishármas maximálisan $\log_2 61 = 5{,}93$ bit információt tartalmazhat, egy bázis pedig 5,93 / 3 = 1,98 bitet. Így a teljes modell-DNS-re $(5{,}93/3) \cdot 4 \cdot 10^6 = 7{,}91 \cdot 10^6$ bit adódik.

A valóságban a DNS nem hordoz ennyi információt, több okból sem. Egyrészt az egyes bázisok valószínűsége nem ugyanakkora (bár gyakran közel egyforma arányban fordulnak elő), másrészt az egymáshoz közeli bázisok valamennyire korreláltak. Ilyen jellegű vizsgálatokkal foglalkozik például *Schmitt és Herzel* [6]. Úgy találják, hogy ezek a hatások nem jelentősek, vagyis a genetikai kód közel optimális: nem tömöríthető számottevően tovább. Másként megfogalmazva, a bázisok sorrendje ránézésre véletlenszerű, nem fedezhető fel bennük rendezett struktúra, azon kívül, hogy elég sok az ismétlődő szakasz. Az élesztőgomba III-as kromoszómáját vizsgálva, a DNS $\boldsymbol{H}$ betűnkénti entrópiájának meghatározásához a H_b/b mennyiségeket becsülik, a $b = 15$ blokkméretből $\boldsymbol{H} \approx 1{,}9$ bit adódik. Az ismétlődő szakaszok hatását az Epstein-Barr vírus genomján vizsgálják: az eredeti szekvenciából $\boldsymbol{H} \approx 1{,}58$ bit a betűnkénti entrópia, míg az összes ismétlődő szakasz (melyek a teljes genom mintegy 25%-át teszik ki) eltávolítása után $\boldsymbol{H} \approx 1{,}95$ bit adódott.

Sokkal jelentősebb tényező, hogy a DNS-nek csak egyes szakaszai kódolnak fehérjéket, a maradék – mélyebb ismeretek hiányában – „hulladéknak” tekinthető. Részben ez magyarázza azt a paradoxont, hogy az élőlények látszólagos bonyolultsága és genomjuk mérete között a kapcsolat nem lineáris. *Adami* [7] információelméleti módszerekkel kísérli meg elkülöníteni egymástól a kódoló és nem kódoló szakaszokat. Érvelése szerint a kódoló szakaszok tekinthetők csak információnak, a nem kódoló szakaszok „csak entrópia”. Hasonlítsuk össze egy egyensúlyban lévő populáció egyedeinek genomját: a DNS egy adott lokusza akkor kódol valami lényegeset az élőlény környezetéről, ha minden egyedben ugyanaz a bázis áll ezen a helyen. Ez ugyanis azt jelenti, hogy a lokusz mutációja életképtelenné tenné az utódot. Ha viszont a lokuszon minden bázis egyforma gyakorisággal fordul elő, akkor az nem tartalmaz a túléléshez szükséges információt az élőlényről. Ezzel a módszerrel beazonosíthatók a DNS kódoló szakaszai, és pontosabban becsülhető a genom által hordozott információ mennyisége.

Vizsgáljuk meg röviden, mit mond az [1] közlemény a biokémiai információ hierarchiájáról. A hivatkozott cikk 4. táblázata azt mutatja be, hogy egy négy bázis hosszúságú RNS molekula különböző reprezentációi mennyi információt tartalmaznak. Ismét a Hartley-entrópiával számol, azaz az összes lehetőség számának kettes alapú logaritmusával. Láttuk azonban, hogy a Hartley-entrópia nem függ attól, hogy a lehetőségeket hogyan reprezentáljuk. Hogyan lehetséges mégis, hogy a különböző reprezentációkra más-más információmennyiség adódik? Úgy, hogy a reprezentációk egyre bonyolultabbak: mind a kódábécé elemszáma, mind az RNS-t leíró sorozatok hosszúsága egyre nagyobb. A táblázat pedig azt tartalmazza, hogy ha *az adott ábécéből az összes ilyen hosszúságú sorozatot* vennénk (nem csak az RNS-eket leírókat), akkor mennyi lenne az entrópia. Formálisan, ha m elemű ábécéből n hosszúságú sorozatokat veszünk, akkor a Hartley-entrópia $n \log_2 m$, tehát akár a sorozatok hosszát növeljük, akár a felhasznált ábécét bővítjük, az entrópia növekedni fog.

Az RNS molekulák alapszerkezete rögzített, így egy RNS molekulát teljesen meghatároz a benne előforduló A, C, G, U molekulák sorrendje, feltéve, hogy biokémiai tudásunk elég alapos, vagyis ismerjük mind az alapszerkezetet, mind az A, C, G, U molekulák szerkezetét,

és ezek kapcsolódási formáját. Természetesen használhatjuk az A – 00, C – 10, G – 01, U – 11 bináris kódot, matematikai szempontból – és így a Hartley-entrópia tekintetében – a kétféle kódolás között nincs különbség. Értelmezhetőségi szempontból az ember számára az A, C, G, U kódolás könnyebben átlátható (míg a számítógép talán a bináris kódot „preferálná"). Ha rövidke RNS molekulánk szerkezetét részletesebben tüntetnénk fel, akkor bővebb ábécét, és hosszabb sorozatokat kellene használnunk (bár ezekben az esetekben valójában nem beszélhetünk sorozatokról, hiszen a szerkezeti képletekben elágazások is szerepelnek). A szerkezeti képlet 421 bit információmennyisége azt fejezi ki, hogy ugyanezekből az atomokból rengeteg különböző (bár nem okvetlenül stabilis) ugyanekkora molekula építhető fel (pontosan $2^{421} \approx 7^{150}$ darab, feltéve hogy a molekulák lineárisak, és a betűk tetszőleges sorozata értelmes molekulát ad). Érdekes kérdés, hogy ebből az öt atomból valójában hány olyan molekula építhető fel, melyben az atomok és a kötések együttes száma éppen 150.

Leírás típusa	**ábécé**	**ábécé elemszáma (m)**	**leírás hossza (n)**	**Hartley-entrópia ($n \log_2 m$)**
bázissorrend	A, C, G, U	4	4	8
bázissorrend bináris kóddal	0,1	2	8	8
vázlatos szerkezeti képlet	A, C, G, U, Cu, F, –	7	24	67
szerkezeti képlet	C, H, O, N, P, –, =	7	150	421

1. táblázat. Négy nukleotidból álló RNS-molekula leírásai négyféle ábécével. Mindegyik ábécére kiszámoltuk, hogy az RNS leírásával megegyező hosszúságú sorozatoknak mennyi a Hartley-entrópiája.

Záró gondolatok

Az információelméletet – a kibernetikával együtt – Shannon alapvető cikkének megjelenése után szinte azonnal üdvözölték az élőlények működésével és evolúciójával foglalkozó kutatók, lásd pl. [8]. Visszatekintve, úgy tűnik, hogy túlzott reményeket fűztek ennek az elméletnek az alkalmazhatóságához. Továbbá Adami és szerzőtársai [9] szerint „Nem újkeletű a törekvés, hogy az információelméletet az evolúció és a szekvenciák információtartalmának megértéséhez segítségül hívjuk. Sajnos azonban számos korai próbálkozás a képet inkább összezavarja, mintsem tisztázná, és gyakran az információ fogalmának téves értelmezésével homályosítja el." Mindez az információelméleti módszerek biológiai alkalmazásának visszaszorulásához vezetett. Ennek ellenére a shannoni elméletnek van létjogosultsága olyan kérdések kutatásában, mint a polimorfizmusok azonosítása, a fehérjék másodlagos térszerkezetének előrejelzése, vagy új gyógyszerek tervezése [7].

Irodalom

[1] Evva Ferenc: A biokémiai információ hierarchiája, *Magyar Kémikusok Lapja,* LXIV. évf. 5. szám, pp. 77–83, 2009.

[2] C. E. Shannon: A mathematical theory of communication, *Bell System Technical Journal* **27** (1948), 379–423 and 623–656.

[3] W. R. Ashby: *Principles of the Self-Organizing System*, In: H. V. Foster and G. W. Zopf (eds.): *Principles of Self-Organization.* Pergamon Press, Oxford, 1962.

[4] R. L. V. Hartley: Transmission of Information, *Bell System Technical Journal* (1928), 535–563.

[5] R. Solomonoff: A Formal Theory of Inductive Inference, *Information and Control* **7** (1964), 1–22 and 224–254.

[6] A. O. Schmitt and H. Herzel: Estimating the entropy of DNA sequences, *J. theor. Biol.* **1888** (1997), 369–377.

[7] C. Adami: Information theory in molecular biology, *Physics of Life Reviews* **1** (2004), 3–22.

[8] H. Quastler (ed.): *Information Theory in Biology.* Urbana: Univ. Of Illinois Press, 1953.

[9] C. Adami, C. Ofria, and T. C. Collier: Evolution of biological complexity, *Proc. Nat. Acad. Sci. USA* **97** (2000), 4463–4468.