Braun Tibor
ELTE TTK Kémiai Intézet, MTA Könyvtár

Adat-, információ- és ismerettárolás az örökkévalóságnak

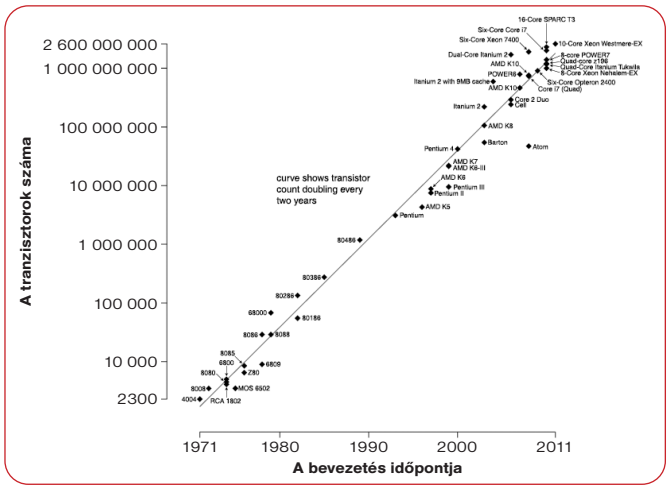
Remények kvarcban és DNS-ben

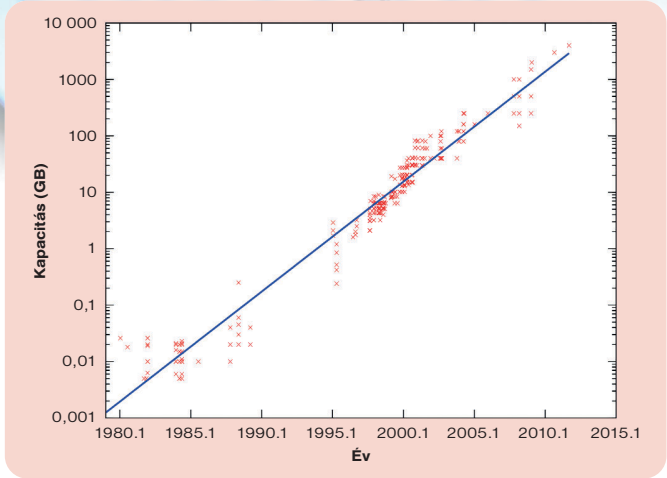
Előszó

Ahhoz, hogy írásunk igazi mondanivalójához eljussunk, induljunk el egy talán primitíven egyszerűsített úton, amelyen feltételezzük, hogy az emberi agyműködés, gondolkodás, illetve tevékenység eredményeképpen adatok keletkeznek, amelyek bizonyos folyamatokon keresztül információkká válnak, majd tudássá lényegülnek. Ezek a folyamatok az emberi agyban játszódnak le, s az adatok ott is maradnak mindaddig, amíg az egyéni agygyal valami végzetes, például halál nem történik, vagy amíg a tárolás érdekében, illetve későbbi felhasználáshoz azokat valahol, valahogyan, valamiben nem rögzítik. Bár az adat-, információ-, illetve ismeretrögzítésnek az emberiség történetében többféle változata alakult ki, például homokban (l. Arisztotelész), agyag- vagy kőtáblákon (asszírok, egyiptomiak), fonalszálakon (inka csomóírás), a mindmáig leginkább elfogadott és használt rögzítési lehetőségek a papírra történő nyomtatás (Gutenberg) és a szilíciumra épülő digitális elektronika.

A fentiekkel kapcsolatos első kérdés az, hogy az említett hordozókra rögzítettek mennyire időtállóak. Az már jó ideje kiderült, hogy valószínű, hogy a papírra nyomtatott szövegek, amennyiben nincsenek agresszív közegeknek kitéve (tűz, víz, nedvesség, vegyszerek), több száz, sőt valószínűleg ezer évnél is tovább fenntarthatók, illetve használhatók. Ezzel szemben a főként műanyag hordozók hordozókon alapuló digitális elektronikai rögzítés (mágnesszalagok, CD- és DVD-rögzítők) évek alatt degradálódik, és a rajtuk rögzített adatok viszonylag gyakran újrafelvételre szorulnak.

1. ábra. A mikroprocesszorokba sűríthető tranzisztorok száma 1971 és 2011 között, (Moore törvénye)

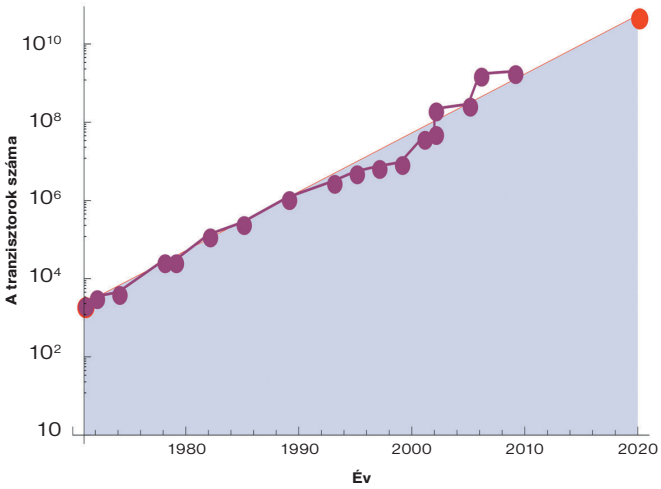




2. ábra. A merevlemezek kapacitásának időbeli változása

A megoldáskeresés kezdeti stádiumában a hosszú időre szóló elektronikus rögzítést az egyre nagyobb kapacitású szilíciumalapú mikroprocesszorokban, illetve az ezeket működtető számítógépekben látták, és ezek alapján egyre nagyobb méretű adatbázisok, adattárolók épültek bízva a Moore törvénye [1] által prognosztizált másfél éves mikroprocesszor- és merevlemez-kapacitások növekedési sebességében (**1.** és **2. ábra**). Bár a jelenlegi tranzisztortechnológiai határok pesszimista szemmel tekintenek a mikroprocesszor-kapacitások Moore törvénye alapján várható növekedésére, optimista futurológusok ezt, legalábbis 2020-ig, lehetségesnek látják (**3. ábra**) [2].

A tartósságon kívül – az egyre nagyobb mennyiségű adattárolásra képes számítógépek építésének folyamatosan növekvő hely-



3. ábra. Moore törvényének kiterjesztése

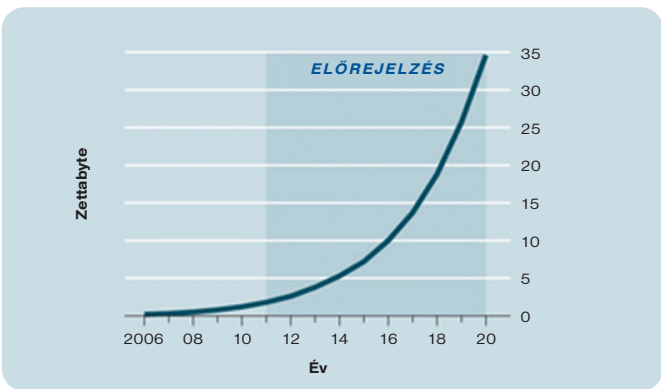
igénye mellett – óhatatlanul előtérbe kerül a szilíciumalapú félvezetők hőtermelési problémája. Ugyanis a sűrűbb, erőteljesebb chipekből épített szerverek melegedése az energia, illetve a hő és a hűtés révén Moore törvényével fordított arányú költségnövekedéshez vezet.

Ennek következtében például a költség/watt tényezők kérdése jelentősen előtérbe került világszerte ott, ahol az adattárolási kapacitás növelése felmerültek, olyannyira, hogy néhol abszurdnak tűnő gondolatok is születtek, mint nagy adattároló központoknak a sarkkör vidékére való telepítése.

Világméretű adatnövekedés

Ahhoz, hogy kissé jobban átláthassuk a címben felvetett témát, nem tudjuk elkerülni a mennyiségek és méretek kérdését [2]. Megemlítjük, hogy időszámításunk kezdete óta 2003-ig az emberiség öt millárd GB ($5{\cdot}5^{15}$ byte) információt hozott létre. Amennyiben ez nem eléggé lélegzetelállító, akkor a növekedési sebesség érzékeltetésére hozzátehetjük, hogy csupán 2011 óra a fenti mennyiség kétnaponként jön létre és ha marad érzékelésre megfelelő levegő, ez az idő 2013-ben tízpercenkéntre zsugorodott [4]. Más szóval a világon tárolandó adatmennyiség exponenciálisan **(4. ábra)** halad a zettabyte (10^{21} byte), sőt a yottabíte (10^{24} byte) felé. Ezek után nem tűnik feleslegesnek elmerengeni azon, hogy az adattárolás jelenlegi fejlettsége, illetve sebessége révén a tárolás milyen helyigényeket és anyagi költségeket vet fel. Egyetlen

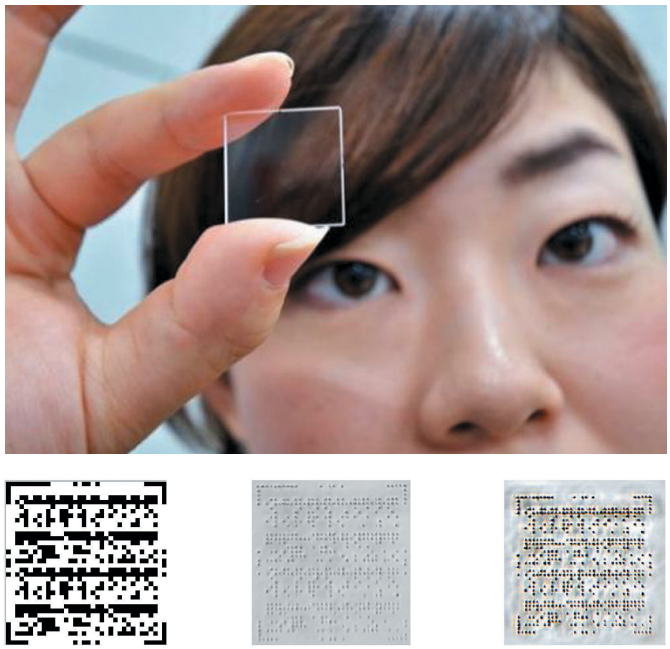
4. ábra. Exponenciális adat- és információnövekedés [7]



példaként megemlítjük, hogy egy IBM-értékelés szerint egy exabyte-nyi (10^{18} byte) orvosi adatmennyiség további feldolgozási célból való mozgatásának költsége könnyen elérheti a 10 millió dollárt. Ez a költség a fájlok igénybevételéhez szükséges jelenleg használt hardvert, szoftvert, energiafelhasználást és munkaerőt foglalja magában.

Adattárolás kvarclapkákban

A közelmúltban a Hitachi, a japán elektronikus óriáscég közzétette, hogy a Kiotói Egyetemmel együttműködve kidolgozták a maratott kvarclapkákon alapuló, millió évekig megmaradó, gyakorlatilag degradáció nélküli adat- és információtárolást [4]. Az elért adatsűrűség 40 MB négyzethüvelykenként, ami valamivel nagyobb, mint a műanyag CD-k körülbelül 35 MB-nyi tárolási kapacitása. A prototípust egy kétszer két centiméter méretű, két milliméter vastagságú maratott kvarclapkán mutatták be. A lapkákba femtoszekundumos (10^{-15} s) lézer által generált fénypulzusokkal négy rétegben pontokat égetnek be. A fénypulzusok bináris, a QR kódhoz hasonuló optikai képet hoznak létre **(5. ábra).** Ezeket a kódolt lapkákat több száz millió évi avulással felérő,



5. ábra. a) a Hitachi-féle kvarclapka b) A lapkára lézerrel égetett QR-kódrendszer http://phys.org/news/2012-09-japan-hitachi.html

két órán át tartó 1000 °C fokú hevítéssel tesztelték. Az optikai mikroszkóppal olvasható kód háborítatlanul megmaradt a lapkákon. Lévén, hogy a kódpontok a kvarclapkák belsejében vannak, azokat felületi erózió nem érinti, mint ahogy savas vagy lúgos vegyületek sem. A lapkák olvasására a Hitachi cég speciális optikai mikroszkópot készített. Mivel a mikroszkóp egy pontrétegre való fókuszálása elhomályosítja a többi réteget, a Hitachi olyan módszert dolgozott ki, ami 15 dB pontossággal szűri ki a homályosodást. Ez gyakorlatilag hiba nélküli adatleolvasást tesz lehetővé. A Hitachi cég olyan gyakorlati megvalósításban gondolkodik a közeljövőben, amiben a kvarclapkákból a Corbis fotógyűjteményhez [6] hasonló archívumokat hoznak létre, amiket nulla Fahrenheit-fokon (– 18 °C) tárolnak.

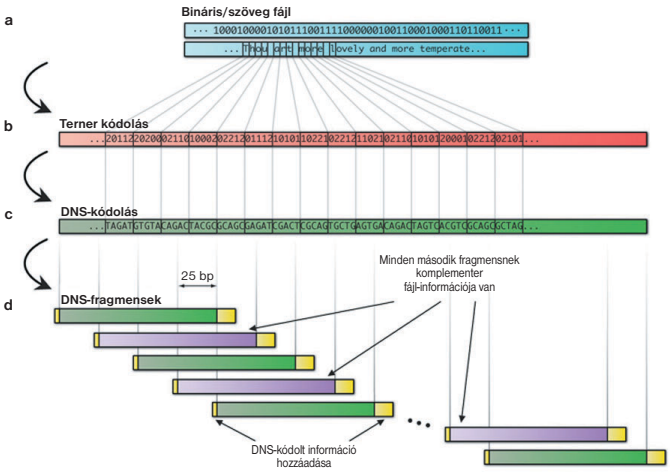
Adattárolás DNS-ben

A DNS alkalmazása adatok tárolására genomok formájában nem tekinthető újnak [8], miután azt teszi minden élő szervezet a Földön [10]. Jelentősnek nevezhető áttörés ezen a téren a közelmúltban történt, amikor angliai, egyesült államokbeli és magyar kutatók közzétették eredményeiket *Towards practical, high-capacity, low-maintenance information storage in synthesized DNA* címmel a *Nature* című folyóiratban [11]. Ezek több szempont szerint is jelentőseknek tűnnek azáltal, hogy 739,3 kB-ra, rekordmennyiségűre sikerült növelni az egyedi kódolt információt. Ezt úgy hozták létre, hogy a jóval szélesebb körű megvalósításokra is lehetőséget nyújtanak. A kidolgozott eljárás nagyjából rögzíteni tudná a világban felhalmozott 3 zettabyte (3×10^{21} byte) információt, jelentős bővítési lehetőséggel. Ezt 2,2 petabyte ($2{,}2\times10^{15}$ byte) sűrűséggel, azaz úgy, hogy elférne egy kisebb teherautó hátlapján, és persze úgy, hogy komoly mértékben csökkenti azokat a másolási hibákat, amelyeknek az eddig javasolt rendszerek ki voltak téve.

Az eljárás újdonsága abban rejlik, ahogy az említett kutatók a fájljaikat a merevlemezekről kémcsövekbe, vagy tintasugaras 2D nyomtatóval üveglapocskákra átkódolják [12]. A DNS négy bázist, az adenozint (A), a timint (T), a citozint(C) és a guanint (G) használja az információ kódolására. A korábbi próbálkozások gyakran a számítógépek által használt bináris 1-gyel és 0-val kódolták ezeket. Például A és C jelentheti a 0-t, illetve G és T az 1-et. A probléma abból adódik, hogy a forráskód 1-es, vagy 0-s szekvenciái egyéni bázis ismétlését okozhatják a DNS-ben (például TTTT). Az ilyen ismétléseket könnyen félreolvassák a DNS-szekvenáló gépek, ezáltal hibákhoz vezethetnek az információ visszaolvasásánál. A megoldást a bináris (kétdimenziós) számítógépes információról a terner (háromdimenziós), azaz három számra, 0, 1 és 2-re való áttérésben, illetve annak az információnak a DNS-be való bekódolásában találták meg. Egy bizonyos kódolási séma attól függ, hogy melyik bázis volt az előző. Amennyiben az előző bázis A volt, akkor a 2-t a T jelzi. De ha az előző bázis a G volt, akkor a 2-t a C képviseli **(1. táblázat).** Hasonló helyettesítési szabályok érvényesek a számok és betűk minden kombinációjára biztosítva azt, hogy az adatokban a hasonló számjegyek (digitek) szekvenciáit ne képviseljék a DNS-ben egymáshoz hasonló számjegyek szekvenciái, ezáltal elkerülve a hibalehetőségeket.

Ezt a kódolást ezután a szintetikus DNS-nél kellett megvalósítani. A legegyszerűbb megközelítés egyetlen hosszú DNS-szál szintetizálása volt minden tárolandó fájlhoz. Sajnos ma még a DNS-szintetizáló berendezések ezt nem tudják megbízhatóan elvégezni. Ezért a fájlokat több ezer egyedi, 117 bázis hosszúságú darabra vágták. Minden darabban 100 nukleotid tükrözi a fájl adatait, a maradék rögzíti azt, hogy a kódolt fájlban hová tartozik egy bizonyos darab. A folyamat tartalmazza a legtöbb számítógéprendszerben található hibajelző „parity bit” DNS-megfelelőjét.

A hibalehetőségek további elkerülésére a forrásfájlokat, mindegyiket kissé különböző átlapoló módon, háromszorosan továbbdarabolták. Ez azért történt, hogy a 100 bázisdarab minden 25 nukleotiddarabból álló negyede is képviselve legyen a DNS három másik darabjában. Amennyiben bármilyen másolási hiba előfordult egy bizonyos darabban, azt össze lehetett vetni a három másik darabbal, és egy többségi helyzet eldönthette, hogy melyik volt a pontos. A darabok visszaolvasása egyszerűen úgy történik, hogy a fragmensekről többszörös másolatokat generálnak sztenderd kémiai reakció útján, majd ezeket a DNS-szekvenálóba táplálják, és ott a fájlokat összeillesztik **(6. ábra).**



6. ábra. A Goldman és munkatársai által feltalált terner DNS-kódolás sematikus bemutatása [11] (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3672958/figure/F1/)

Amikor ezt a sémát tesztelték, csaknem a tervezett szerint működött. Az említett kutatóknak sikerült öt számítógépes fájlt kódolni és dekódolni, ezek között *Martin Luther King* „I have a dream” című MP3-on rögzített beszédének egy részét, illetve *Francis Crick* és *James Watson* 1953-as, a DNS szerkezetét leíró cikkének PDF-változatát. Az egyetlen gubanc az volt, hogy minden elővigyázat ellenére a DNS-cikk két 25 bázisnyi szegmense elszállt. A hibát végül is egyrészt a DNS kémiájának, másrészt a szintetizálásnak tulajdonították.

Mint mindenben, a DNS-ben való adattárolásnak is vannak árnyoldalai. Az egyik az a relatív lassúság, amivel az adatok visszaolvashatók. A kutatóknak két hetét vette igénybe a fent említett öt fájl visszaépítése, bár leírták, hogy jobb berendezéssel egyetlen nap alatt is elvégezhető lett volna. És persze a folyamat tovább gyorsítható lett volna több szekvenáló berendezés igénybevételével. A másik hátrányt a jelenlegi stádiumban a költségek jelentik. Kiszámították, hogy kereskedelmi árakon a DNS-tárolás kódolása körülbelül 12 500 dollárba kerülne tárolt MB-onként. Ehhez jönne MB-onként a dekódolás 200 dollárja. Tehát a jelenleg használt mágnesszalagon való digitális információtárolásnak ez 10^6-szorosába kerülne.

1. táblázat. A háromdigites kódolás

Kódolt digit	Kódolt aminosav: A	C	G	T
0	C	G	T	A
1	G	T	A	C
2	T	A	C	G

De mint a bevezetőben már említettük, a mágnesszalagok degradálódnak, öregszenek és néhány évenként újakkal pótolandók, míg a DNS milliárd évekig olvasható marad, ha hűvös, sötét és száraz helyen tárolják, amint a neandervölgyi gyapjas mamutok maradványaiból kivont DNS-leolvasás bizonyította.

Minél hosszabb időre tárolandó az információ, annál vonzóbbá válik a DNS, azt nem is említve, hogy a DNS-szintézis és -szekvenálás költségei rohamosan csökkennek. Úgy tűnik, hogy körülbelül egy évtized múltán a DNS versenyképessé válhat olyan adatbázisok tárolására, amelyek ötven évre, vagy hosszabb időre megőrzendők.

Utószó

A technológia kutatása óhatatlanul előtte jár a technológia pillanatnyi fejlődési szintjének, de természetesen igyekszik azt utolérni. Ez annál is inkább igaz, ha a fejődés már kopogtat a mindennapok kapuján, de a járulékos szükségletek bizonyos mértékben lemaradtak. Gondolunk itt a kvarclapkákat olvasni tudó gyors és önműködő mikroszkópokra, vagy a DNS-láncokat szintetizáló, illetve szekvenáló még gyorsabb, olcsóbb berendezésekre. Nem kétséges, hogy a lemaradás behozható, sőt ez meg is fog valósulni. Az, hogy ezzel az alapkérdés megnyugtatóan megoldódik-e, vagy akkorra, amikor a lemaradás megteszi az utolsó lépést, az alapkérdés – más irányból szemlélve – megint eltávolodik, ma még nyitott kérdés.

IRODALOM

[1] http://en.wikipedia.org/wiki/Moore%27s_law

[2] http://www.zdnet.com/2020-the-visible-end-of-exponential-innovation-7000019987/ (utolsó hozzáférés: 2013. október 18.)

[3] P.Thibodeau: China's 'big hole' marks scale of supercomputing race. Computerworld, http://www.computerworld.com/s/article/9187702/China_s_big_hole_marks_scale_of_supercomputing_race (utolsó hozzáférés: 2013. október 17.)

[4] P. Gilster: Data Storage: The DNA Option. Centauri Dreams. Imagining and Planning Interstellar Exploration. January 28, 2013, http://www.centauri-dreams.org?p=26262 (utolsó hozzáférés: 2013. október 17.)

[5] M. Sakakura, T. Tochio, Y. Ishiguro, M. Nakabayashi, Y. Shimotsuma, K. Hirao, K. Miura: Dynamics of fracture inside a single crystal induced by a focused femtosecond laser pulse, *Proc. SPIE 8530*, December 4, 2012, DOI: 10.1117/12.976339 in: G. J. Exarhos, V. E. Gruzdev, J. A. Menapace, D. Ristau, M. J. Soilean, (Eds.): *Laser Induced Damage in Optical Materials*, Boulder, Colorado, September 23, 2012.

[6] http://hu.wikipedia.org/wiki/QR_k%C3%B3d (utolsó hozzáférés: 2013. október 18.)

[7] History flushed. Digital archiving, The Economist, April 28th 2012, http://www.economist.com/node/21553410 (utolsó hozzáférés: 2013. október 17.)

[8] E. B. Baum: Building an associative memory vastly larger than the brain. Science (1995) 268, 583.

[9] http://en.wikipedia.org/wiki/DNA_digital_data_storage (utolsó hozzáférés: 2013. október 17.)

[10] http://ghr.nlm.nih.gov/handbook/basics/dna (utolsó hozzáférés: 2013. október 18.)

[11] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, M. LeProust, B. Sipos, E. Birney: Towards practical, high-capacity, low-maintenance information storage in synthesized DNA, *Nature*, (2013) 494, 77. DOI:10.1038/nature11875

[12] DNS – Az adattárolás jövője, http://www.peetandcook.hu/component/k2/item/107-dns-az-adatt%C3%A1rol%C3%A1s-j%C3%B6v%C5%91je (utolsó hozzáférés: 2013. október 17.)

ÖSSZEFOGLALÁS

Braun Tibor: **Adat-, információ- és ismerettárolás az örökkévalóságnak. Remények kvarcban és DNS-ben**

Az adatok hosszú időre való tárolása ma még meg nem oldott feladatot jelent. A dolgozat a jelenleg még kutatás tárgyát képező megoldásokat veszi számba és elemzi.