Az IT szerepe a genomikában

FALUS ANDRÁS
Semmelweis Egyetem
falus.andras@med.semmelweis-univ.hu

Kulcsszavak: genetika, genom, genomika, bioinformatika

A molekuláris biológia robbanásszerű fejlődése az elmúlt hetven évben lehetővé tette az örökítő anyag szerkezetének és működésének megismerését a teljes földi élővilágban. A technikai fejlődés által rendelkezésre álló gigantikus adatmennyiség feldolgozása és a működési mechanizmusok feltárása nyomán az informatika elengedhetetlen részese lett a genomikai kutatásnak. A cikk a genomikai forradalom hátterével, egyes alkalmazási területeivel és perspektívájával foglalkozik.

1. A genetika és genomika tudománya

Az emberi intellektus egyik legnagyobb közös teljesítménye a humán és nagyszámú más élőlény örökítő anyagának molekuláris szintű megismerése. A páratlan nemzetközi összefogással, 1989–2003 között megvalósult humán genom programról (HUGO – Human Genome Organization) túlzás nélkül állíthatjuk, hogy elkezdődött vele a biológia „írásbelisége".

Az emberi szervezetben mintegy száz billió (10^{14}) sejt található. Minden egyes sejtünk sejtmagjában 2x23 kromoszóma (ivarsejtekben a fele), 2x3,2 milliárd (10^9) négyféle *nukleotidbázis* (ezek: adenin-A, guanin-G, citozin-C és timin-T) található, ez a dupla helikális szerkezetben kb. 2x2 (4) méter DNS-t (dezoxiribonukleinsavat) jelent. Az RNS (ribonukleinsav) timin helyett uracilt tartalmaz. A nukleotidbázisok lineáris sorrendje képezi a szüleinktől örökölt „biológiai hardvert". Az élet során különböző hatásokra természetesen megváltozhatnak a nukleotidok (csere, kiesés, beékelődés, átrendeződés), ezeket a változásokat *mutációknak* nevezzük.

Az egyes génekkel a *genetika,* az összessel (beleértve azok kölcsönhatásával is) pedig a *genomika* foglalkozik. A genetika legfontosabb felfedezéseinek (a DNS mint örökítőanyag azonosítása, az öröklődés törvényeinek felismerése, a DNS szerkezetének leírása) sorába illik óriási továbblépésként a teljes human örökítőanyag (genom) szekvenciájának megállapítása.

A szekvenálás, tehát a nukleotidbázisok sorrendje legközvetlenebb eredménye az összes – fehérjét kódoló – mintegy 23-25 ezer gén azonosítása. Az emberi gének viszonylag csekély száma (mely nagyságrendileg hasonló a fonalféregben találtakhoz!) rávilágított arra, hogy a biológiai fenotípus (tehát a valóságos megjelenés) komplexitását nem a génkészlet nagysága, hanem magukban a gének variánsaiban rejlő egyedi sokféleség (diverzitás), a kapcsolati gén- és géntermék-hálózatok szövevénye, valamint a gének megszólalására ható epigenetikai hatások sokasága határozza meg. A gének katalogizálásán túlmenően a genomszekvencia megadja a gének pontos helyét és sorrendjét is a kromoszómákon. Ez az egyszerű információ óriási jelentőségű a genetikában, mert lehetővé teszi azt, hogy egy kromoszóma szakaszhoz kapcsolt („térképezett") tulajdonsághoz vagy betegséghez gének módosulásait, variációit rendelhessünk hozzá.

A genomot tekintve csillagászati méretekről van szó, hiszen ha az összes emberi sejttel számolunk, az emberi szervezet DNS-hossza mintegy 140-szerese a Föld-Nap távolságnak.

Az emberiség DNS szinten is nagyon egységes, a raszszizmus minden álságos biológiai alátámasztása nemcsak hogy morálisan elfogadhatatlan, hanem biológiailag, tudományosan is hamis. Az egyes etnikumok között néhány tizedszázalékos eltérés van a genom szintjén.

2. A genomikai korszak fő „inputjai"

A nyolcvanas évektől kezdődő szinte példanélküli tudományos robbanás főként három forrásból táplálkozott:

1. Hatalmas mértékben felgyorsult a nanobiotechnológián alapuló, nagy áteresztő képességű, ún. „high-throughput" metodikák fejlesztése (nukleotid szekvencia-meghatározás, microarray technológia, teljes genom vizsgálatok tekintetében); egyre nagyobb kutatási teljesítményt egyre olcsóbban lehetett elérni. A közelmúltból külön kiemelendők az újgenerációs, szinte teljesen automatizált szekvenálási, valamint a valódi, precíz génterápiára reményt nyújtó génszerkesztési eljárások hatása.
2. Egyre több és egyre teljesebb szabadon elérhető adatbázis vált hozzáférhetővé a kutatók számára. Létrejött a térbeli és időbeli korlátokat virtuálissá tevő *„in silico"* (komputer előtti) kutatás lehetősége. Ez egyben a tudomány rendkívül széleskörű demokratizálódásával járt, hiszen bárki a világon könnyen és a legtöbb esetben ingyen felkeresheti ezeket az adatbázisokat interneten. Ezt követően saját kutatólaboratórium (*„nedves labor"*) nélkül is a meglevő

adatok új megközelítésével, csoportosításával, csupán a számítógép mellett dolgozva, önálló és eredeti tudományos felfedezéseket tehet.

3. Szükségszerűen kiteljesednek a bioinformatikai elemzések a nagy elemszámú biológiai rendszerek adattengerének elemzésére is. Napjainkban útvonal- és génhálózat-analízisek, valamint az ennek megfelelő szoftverek sokasága jelent és jelenik meg.

Az *1. ábrán* az emberi májrák egy génhálózati kapcsolatrendszerét mutatjuk be részletezés nélkül (csak a hálózati komplexitást szeretnénk demonstrálni). Az egyes szimbólumok a tumor működésére ható géneket szimbolizálják. A háromszög alakú jelek a genom stabilitását biztosító géneket, a köralakú szimbólumok pedig a programozott sejthalál funkcióban résztvevő géneket jelzik.

A hagyományosabb, úgynevezett „frekventista" analízisek mellé beléptek a nagy halmazokat kezelő matematikai-statisztikai eljárások. Ezek között például a BNBMLA (Bayesian multilevel analysis) véletlen változók közötti kapcsolatok valószínűségének eloszlását mutatja.

Nem véletlen, hogy ma már a molekuláris- és genomszintű vizsgálatok anyagi feltételei közül a szuperszámítógépek, a folyamatosan megújított szoftverek és az azokat fejlesztő, jól felkészített informatikusok (bioinformatikusok) tudása minősíthető az egyik legkeresettebb (és jól fizetett) hivatásnak.

3. A genomika és az informatika kapcsolata

A modern genetika tehát ma már elválaszthatatlanul kapcsolódik az informatika tudományához. A fentebb említett humán genom program lehetővé tette e gigantikus információ „elolvasását". A program 13-14 évében, rendkívüli nemzetközi együttműködéssel leírták a genomot alkotó, mintegy 3,2 milliárd építőelem (nukleotidbázisok: A, C, G és T) lineáris sorrendjét. A két nagy konkurrens, az államilag támogatott HUGO, illetve a Celera (privát cégből kinőtt magánvállalkozás) természetesen csak kevés egyedi genomot tudott „elolvasni", ennek megfelelően messze nem volt világos, hogy mely genetikai „szavak és betűk" találhatóak meg minden emberben, és melyek valóban egyediek.

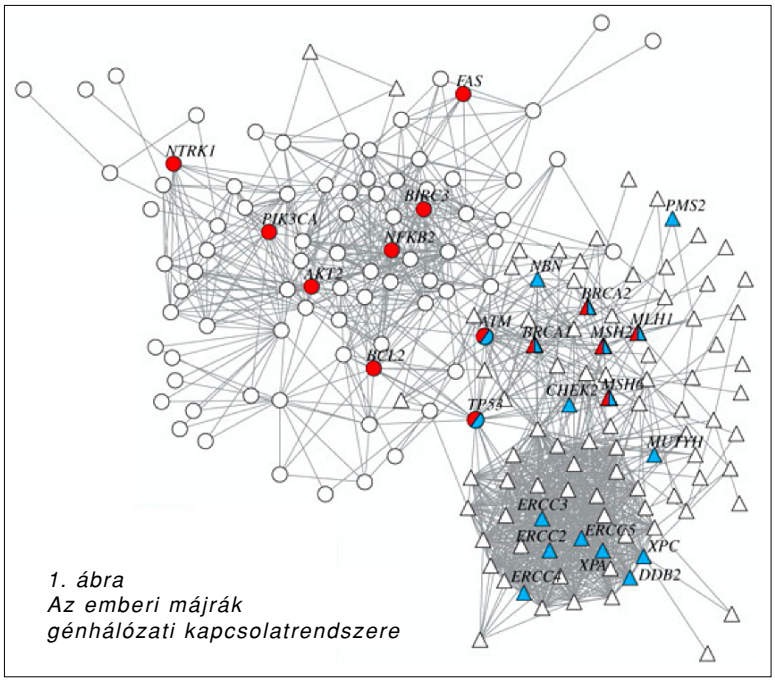
Egy emberből átlagosan 20 fehérjekódoló gén teljesen hiányzik, azaz ebből a szempontból „génkiütöttnek" tekinthető. Ezek általában olyan gének, melyek hiánya nem okoz evolúciós hátrányt a ma élő embernek. Ilyenek például egyes szagreceptorok hiányai. Vannak viszont olyan génhiányok, amelyek kisebb hátrányt, vagy előnyt jelenthetnek hordozójuknak.

A humán genom 45%-a ismétlődő szekvenciákból áll. Ezek közül sok a *transzpozon,* azaz ugráló gén, amelyek viszont akár 40 millió év óta is inaktívak. A leggyakoribb ismétlődő szekvenciát Alu-nak hívják, mely a teljes genomunk 10,6%-át foglalja el.

Több száz génünk származik baktériumokból horizontális gén-transzferből. A pericentromerikus és a subtelomerikus régiókban nagy szakaszok ismétlődnek.

Jelenleg az *imprintált* gének számát 150 körülire becsülik (genetikai imprinting: az eltérő apai és az anyai gének kifejeződése), amelyek közül vagy csak az anyai (56%), vagy csak az apai (44%) aktív, de a pontos számok vitatottak. Ha valami oknál fogva ebben a rendszerben hiba következik be, például ha mindkét gén aktív, ez többféle súlyos betegségekhez is vezethet.

A legújabb definíció szerint a *paralógok* ugyanabban a fajban található közös gének, míg az *ortológok* hasonló, különböző fajokban levő gének. A paralóg gének génduplikáció eredményei működnek; van intronos (unprocessed) vagy intronnélküli (processed) változat, funkciója lehet ugyanaz, vagy hasonló, de más is, mint az eredeti génnek. Az úgynevezett „processzált" paralóg úgy keletkezik, hogy a génből átíródott mRNS-ből splicing útján kivágódnak az intronok, majd reverz transzkripció után visszamásolódnak a genomba. Mivel a szelekciós nyomás a duplikálódott génen kisebb vagy hiányozhat, szabadon mutálódhat, így nyerve új funkciókat.



1. ábra
Az emberi májrák génhálózati kapcsolatrendszere

A humán genomszekvencia sikeres leírása, első „munkapéldánya"(„draft"-ja) pontatlanságai ellenére is vitathatatlan mérföldkő volt a genetikában, hiszen a genomszekvencia a genetika olyan alapdokumentummá vált, ami nélkül a genetikai tudományok további fejlődése elképzelhetetlen volt. Olyan ez, mint egy könyv szövege, betű- és szóhalmaza, ami szükséges – de nem elégséges – feltétel a „szöveg" megértéséhez. Önmagában ezzel az „írásjeltömeggel" még nem tudunk mit kezdeni, a nyelv, a biológiai „nyelvtan" ismerete nélkül csak értelmetlen ákom-bákomnak látjuk.

Ma már az úgynevezett „posztgenomikus” korban élünk, a lexikális megismerésen túl a működés, szabályozás és a gének funkcióinak feltárása, az „annotáció” zajlik. Megtudtuk, hogy az örökítő anyag óriási elemszámú hálózatokban működik. A teljes rendszer áttekintését célzó megközelítésre szolgál a *rendszerbiológia* vagy rendszerszemléletű biológia (systems biology) elnevezés, amely egy teljesen új „csapat-függő” világot nyitott meg a kutatók, orvosok, biotechnológusok és matematikusok számára.

Nyilvánvaló, hogy tudásunk validálásához még sokkal több ember genomszekvenciájának megismerésére lesz szükség.

2012-ben fejeződött be az ún. „1000 genom projekt”, ennek alapján jött létre az ENCODE, ami egy genetikai enciklopédiának felel meg. Kínai genetikusok közeli célul tűzték ki több millió ember teljes genomjának elolvasását. A viharosan fejlődő módszerek, például az újgenerációs szekvenálási eljárások és a rohamosan csökkenő költségek folytán valószínű, hogy ez a cél pár éven belül meg fog valósulni.

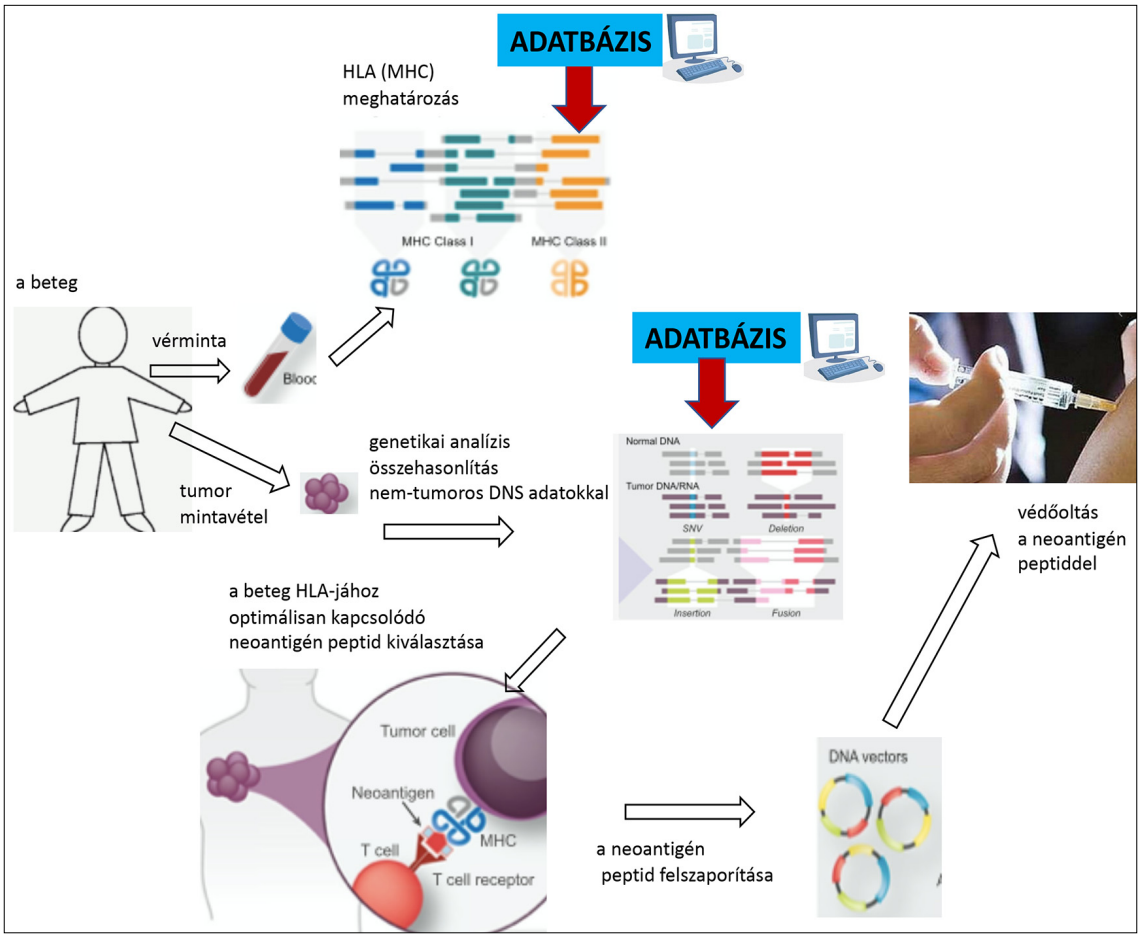
A továbbiakban meg kell tudnunk mondani minden egyes variánsról, hogy hozzájárul-e a betegséghez, vagy például egy adott gyógyszer lebontásának kinetikájához, s ha igen, milyen mértékben. Ennek megállapítása igen nehéz feladatnak ígérkezik, tekintve, hogy a betegségeket okozó variánsok száma valószínűleg igen nagy és a legtöbb etnikumban, sőt egyes emberekben is különböző.

Mindazonáltal ennek a genetikai információnak a birtokában prediktív módon megbecsülhető lesz majd a betegségek kialakulásának genetikai kockázata, még azok bekövetkezte előtt.

Az informatikai analízisek egyre inkább a mesterséges intelligenciák alapvető felhasználása felé mutatnak. Az informatikai analízis sémáját a rákkutatás egy példáján keresztül mutatjuk be.

A tumorokra jellemző neoantigének kimutatása és aminosav sorrendjének meghatározása az egyik első kulcseleme a rákkutatásnak. Ma már ez a vizsgálat bioinformatikai jellemzések sorozatán keresztül valósul meg. A neoantigén jellemzésére és a tumorvakcina bioinformatikai előállítására szolgáló átfogó munkafolyamat főbb elemzési lépéseit egyszerűsített formában a *2. ábrán* mutatjuk be.

Először betegek örökölt immungenetikai sajátosságait, a *fő hisztokompatibilitási fehérjék* – MHC (emberben: *humán leukocita antigén* – HLA) típusait határozzák meg, felhasználva a nagy adatbázisokban talált információkat. Ezután a tumor genetikai analízise következik, variánsokat keresnek a betegből izolált (biopszia, műtéti minta) tumorszövetben (például nukleotid cseréket, egyes szakaszok kiesését, beépülését vagy összekapcsolódását). Ezt követően az MHC-fehérjékhez kapcsolódó neoantigén-peptidek közül informatikai („in silico”) predikciót hajtanak végre az interneten hozzáférhető adatbázisok felhasználásával, azaz a beteg HLA-molekuláira „illesztve” tervezik meg a legjobban kapcsolódó neo-



2. ábra
A tumorvakcina bioinformatikai előállítására szolgáló munkafolyamat főbb elemzési lépései

antigén eredetű peptideket. A kiválasztott peptideket ezután megfelelő hordozókkal (pl. vírusok) felszaporítják és vakcinákat állítanak elő. A vakcinák stimulálják a beteg immunrendszerét, és immunológiai védelmet nyújtanak a daganat ellen. A kutatások nyomán egyre hatékonyabb tumorellenes védőoltások előállításával reményteli módon fel lehet venni a harcot a molekulárisan jellemzett daganatok ellen.

4. A „geneticizmus" veszélye

A genetikai/genomikai/epigenetikai „hype" (csinnadratta), a genetika kizárólagos jelentőségének túlhangsúlyozása („geneticizmus") nagy veszélyt is jelenthet, mert a megismerés, a tudásunk és a gyakorlati hasznosíthatóság jelen fázisa kezdetinek tekinthető. A nagy hírveréssel nyilvánosságra hozott ENCODE eredményei pár éve elmaradtak a várakozástól. Megjelent egy elég szkeptikus kifejezés; a *hiányzó örökletesség* (missing heritability). Ez persze nem az eredményeket, hanem a túlzóan (de talán érthetően) nagy elvárásokat minősíti.

A genetika/genomika értelmezése markánsan eltávolodik a „sors" fogalmától, ma már e tudományok legfontosabb szavainak egyike a hajlam, amibe a valószínűség fogalmát is bele kell értenünk. A hálózati gondolkodás mellett a külső és belső környezet által létrejövő, reverzibilis epigenetikai hatások figyelembe vételével sokkal jobban helyére kerülnek a genetika és a genomika tudományának óriási és egyben valós eredményei és társadalmi hasznossága.

A szerzőről

FALUS ANDRÁS (szül.1947) biológus, a Semmelweis Egyetem professor emeritusa, a Genetikai-, Sejt és Immunbiológiai Intézet igazgatója 1994–2012 között, Pro Universitate- és Széchenyi- díjas, az MTA és az Academia Europeae rendes tagja. Fő kutatási területe az immungenomika és az epigenetika. Hosszabb ideig dolgozott Odenseben (Dánia), a bostoni Harvard és az Osakai Egyetemen. Tizenkét nemzetközi és magyar tudományos könyvet írt és szerkesztett. Több mint 400 tudományos publikációjára mintegy 14.600 idézetet kapott, h indexe: 55. Negyven PhD hallgatója szerzett fokozatot. Kutatási tevékenysége mellett ismeretterjesztéssel is foglalkozik. Társalapítója az EDUVITAL Nonprofit Egészségnevelési Társaságnak. Nős, három gyermeke és 12 unokája van.

Hivatkozások

[1] *http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml* (2009).
[2] International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature 2001; 409:860–921.
[3] Venter J.C. et al.: The sequence of the Human Genome. Science 2001; 291:1304–1351.
[4] International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. Nature, 2004 Oct 21; 431:931–945.
[5] *http://genomics.xprize.org/*
[6] Rusk, N., Kiermer, V.: Primer: Sequencing – the next generation. Nature Methods 2008; 5:15.
[7] *http://www.genome.gov/10005107* (2009).
[8] Pennisi, E.: 1000 Genomes Project gives new map of Genetic Diversity. Science 2010; 330:574–575.
[9] *http://www.epigenome.org/* (2009).
[10] Redon, R. et al.: Global variation in copy number in the human genome. Nature 2006; 444:444–454.
[11] Armour, J.A.: Copy number variation and antigenic repertoire. Nature Genetics 2009; 41 (12):1263–1264.
[12] Bruder, C.E. et al.: Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am. J. of Human Genetics 2008; 82:763–771.
[13] Pauline C. Ng, et al.: Genetic variation in an individual human exome. PLOS Genetics, 2008 Aug 15; 4 (8):e1000160.
[14] Reich, D. et al.: Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature, 2010 Dec 23; 468 (7327):1053–1060.
[15] Green, R.E. et al.: A draft sequence of the Neandertal genome. Science, 2010 May 7; 328 (5979):710–722.
[16] Reich, D. et al.: Denisova admixture and the first modern human dispersals into southeast Asia and Oceania. Am. J. of Human Genetics, 2011 Oct 7; 89 (4):516–528.
[17] Burbano, H.A. et al.: Targeted investigation of the Neandertal genome by array-based sequence capture. Science, 2010 May 7; 328 (5979):723–725.
[18] Gibbs, W.W.: The unseen genome: gems among the junk, Scientific American 2003; 289 (5):46–53.
[19] The ENCODE Project Consortium: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447:799–816.
[20] Parker, S.C., Hansen, L., Abaan, H.O., Tullius, T.D., Margulies, E.H.: Local DNA topography correlates with functional noncoding regions of the human genome. Science 2009; 324:389–392.
[21] Fire, A., Xu, S., Montgomery, M., Kostas, S., Driver, S., Mello, C.: Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 1998; 391 (6669):806–811.
[22] Swami, M.: RNA world – A new class of small RNAs. Nature Reviews Genetics 2009; 10:425.
[23] Waterston, R.H. et al.: Initial sequencing and comparative analysis of the mouse genome. Nature 2002; 420 (6915):520–562.