**BIOINFORMATIKA**

**Ungvári Ildikó**

**Összefoglalás**

A bioinformatika a biológia és az informatika határterületén formálódott interdiszciplináris tudományterület, ami biológiai és orvosi problémák, kérdések informatikai jellegű vizsgálatával foglalkozik.

A természettudományos kutatásoknak lendületet adó óriási technikai fejlődés okaként és egyben következményeként a bioinformatika pillanatok alatt tört be a kutatók mindennapjaiba, olyannyira, hogy ma már nehéz lenne a biológiáról bioinformatika nélkül gondolkodni. Ez a rövid összefoglaló a bioinformatika alapjaiba, és legfőbb alkalmazási területeibe nyújt betekintést.



Az elmúlt évtizedekben a molekuláris biológiai, genomikai technológiák robbanásszerű fejlődése a biológiai adatok mennyiségének exponenciális növekedéséhez vezetett. Ebben a folyamatban mérföldkőnek tekinthető a Sanger-féle nukleotidszekvenálás automatizálása, illetve ennek folyományaként az 1990 és 2004 között zajló humángenomprogram, amelyek következtében nagyon gyorsan felszínre kerültek az adatok tárolására, rendszerezésére és analízisére irányuló megemelkedett igények. Ráadásul az első genomprogram óta a szekvenálási (genom nukleotidsorrendjét meghatározó), illetve genotipizálási (genetikai eltéréseket felfedő) eljárások a technika rohammértékű fejlődésével arányosan, napról napra egyre olcsóbbá válnak, olyannyira, hogy ma már talán az sem tűnik utópisztikus gondolatnak, hogy megismerjük minden egyes ember genetikai állományát, s ezáltal személyre szabott gyógyszereket, vakcinákat fejlesszünk. Ez a ma is zajló technológiai forradalom a számítógépeken tárolt biológiai információ már-már kezelhetetlen mértékű felhalmozódását eredményezi.

A tudomány fejlődésével az is nyilvánvalóvá vált, hogy az élettani működések vagy a populáció nagy részét érintő komplex, multifaktoriális betegségek nem értelmezhetők egy-egy gén vagy fehérje működése alapján, hanem nagyon sok, esetenként több száz gyenge (genetikai, epigenetikai vagy környezeti) kölcsönhatás bonyolult hálózatokat alkotva játszik szerepet azok

létrejöttében. A hagyományos, egy gén – egy betegség megközelítésünket a 20. század végén rendszerszemléletre (systems biology) váltottuk fel, és az emberi agy képességeit meghaladó, összetett hálózatokban kezdtünk el gondolkodni.

Ezt az elképesztő technológiai és tudományos fellendülést nyilvánvalóan szorosan kellett kísérnie a bioinformatika fejlődésének, és ma a biológiai és orvostudományok már elképzelhetetlenek lennének bioinformatikai támogatás nélkül. Napjainkban az elméleti alapkutatással foglalkozó kutató és a klinikus sokféle genomikai, proteomikai eljárást alkalmaz, folyamatos kapcsolatot tart az informatikusokkal, akik segítségével biobankokat hoz létre, mesterséges modelleket alkot, és molekuláris hálózati útvonalanalízist alkalmaz.

**A bioinformatika feladata:**

– molekuláris biológiai, genetikai és biokémiai adatbázisok létrehozása, illetve az azokban tárolt információk kinyerésére és analízisére szolgáló eszközök, módszerek fejlesztése;
– a tudományos kísérletek tervezésének elősegítése;
– a tudományos kísérletek eredményeinek vizsgálatára alkalmas statisztikai megoldások fejlesztése, illetve támogatás nyújtása az eredmények értelmezésében;
– döntéstámogatás a gyógyászatban (diagnózisban, kezelésben és előrejelzésben egyaránt).

**Biológiai adatbázisok**

A biológiai adatbázisok a tudományos kísérletek, a publikált tudományos irodalom és a bioinformatikai elemzések eredményeinek legtöbbször ingyenesen elérhető tárházai. Az összetettebb, ún. relációs adatbázisokban nemcsak kereshetők a tárolt adatok, hanem egyszerűen használható szoftverek segítségével különböző elemzéseket is elvégezhetünk rajtuk. A teljesség igénye nélkül beszélhetünk nukleotid-, fehérje-, útvonal- vagy akár betegség-adatbázisokról, mégpedig az organizmusok széles palettáján. Így az eukarióta-adatbázisokon kívül számtalan vírus- vagy baktérium-adatbázissal is találkozhatunk.

A Nucleic Acids Research folyóiratban 2004 óta évenként megjelenik egy adatbázisokról szóló kötet (http://www.oxfordjournals.org/nar/database/c). A 2010-es kiadásban például 1230 kiemelt jelentőségű adatbázis gondosan összeállított listáját és rövid leírását találjuk meg. A főbb adatbázisok között találjuk az NCBI (National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/) számos adatbázisát, amelyekből a központi szerepű Entrez (http://www.ncbi.nlm.nih.gov/Entrez/) keresőrendszer segítségével integrált, azaz különböző adatbázisokból származó információt nyerhetünk ki.

Például egy számunkra érdekes gén (az oxigén szállításáért felelős hemoglobinmolekula béta-alegységét kódoló gén, HBB) nevét beírva a keresőfelületre az NCBI különböző adatbázisainak találati listáját kapjuk meg, amelyek egymással is keresztkapcsolatban állnak. A különböző nukleotid-

adatbázisokban megismerhetjük a gén nukleotidszekvenciáját, kromoszomális elhelyezkedését, de információt kapunk a génben található variációkról, azok hemoglobinmolekulában betöltött szerepéről és a variációk következtében kialakuló betegségekről is (pl. vérszegénységek). A fehérje-adatbázisokban megtaláljuk a génről átíródó fehérje aminosavsorrendjét, 3D térszerkezetét, valamint a kisebb szerkezeti motívumokat, alegységeket. Emellett képet kaphatunk az eukarióta organizmusokban fellelhető homológiákról, vagyis olyan hemoglobin-B-génterületekről, amelyek például a csimpánzban, egérben és emberben is nagyfokú szekvenciahasonlóságot mutatnak.

Az NCBI PubMed adatbázisán keresztül pedig elérhetjük a hemoglobinmolekulával kapcsolatos tudományos szakirodalmat.

Az NCBI adatbázisait külön-külön linken is elérhetjük, például:

– Genbank: DNS- és aminosavszekvencia-adatbázis, 2009-ben több mint 100 milliárd bázispárt, illetve több mint 100 millió egyedi szekvenciát tartalmazott:
– http://www.ncbi.nlm.nih.gov/genbank/

A géneket/szekvenciákat megtekinthetjük szöveges formában vagy a Map Viewer alkalmazás segítségével térképes megjelenítésben is, ahol nemcsak a vizsgált génről, de az adott genomterületen elhelyezkedő további génekről, illetve azok egymáshoz viszonyított kromoszomális elhelyezkedéséről és távolságáról is információt kaphatunk.

– Az egynukleotidos variációkat (SNP) tartalmazó dbSNP adatbázis: http://www.ncbi.nlm.nih.gov/projects/SNP/
– A több mint 20 millió orvosi/orvosbiológiai cikket tartalmazó PubMed adatbázis: http://www.ncbi.nlm.nih.gov/pubmed
– Az összes ismert, genetikai okokra visszavezethető humán megbetegedés internetes katalógusa az OMIM (Online Mendelian Inheritance in Man), részletes szakirodalmi áttekintéssel a klinikai paramétereket és a meghatározó genomterületeket illetően: http://www.ncbi.nlm.nih.gov/omim

Említésre méltó még a nem az NCBI által fenntartott GO (Gene Ontology, http://www.geneontology.org/) adatbázis, amely tájékoztatást nyújt a génekről kifejeződő fehérjék molekuláris funkciójáról, valamint a különböző biológiai folyamatokban való részvételéről. Ezen kívül megtudhatjuk azt is, hogy az adott protein (ha például szerkezeti elemről van szó) milyen celluláris alkotórész kialakításában vesz részt. Például a hemoglobin esetében a fehérje molekuláris funkciói közt találjuk többek között az oxigén-, fehérje-, illetve fémionkötést, biológiai folyamatai közt az oxigénszállítást, a vérnyomás-szabályozást, a hozzá tartozó celluláris komponens pedig a hemoglobinkomplex.

**Főbb bioinformatikai módszerek**

***Szekvenciaanalízis***

Az új technológiáknak köszönhetően ma már több mint hatezer különböző organizmus teljes genomiális szekvenciáját ismerjük. Ezen szekvenciák bioinformatikai elemzésével, az ún. genomannotáció során fehérje- vagy RNS-kódoló géneket, szabályozó régiókat, szerkezeti motívumokat vagy az egyes betegségekre jellemző repetitív szekvenciákat ismerhetünk meg.

A szekvenciaanalízis egyik metódusa a homológiakeresés, amelynek során például egy új, ismeretlen funkciójú génhez/fehérjéhez tartozó szekvenciához próbálunk hasonlót találni a már ismert funkciójú gének/fehérjék szekvenciái között, majd a talált hasonló szekvenciák funkciójának ismeretében feltételezést teszünk a vizsgált gén vagy fehérje funkcióját illetően.

Számítógépes szekvenciaillesztő programok (mint amilyen az NCBI BLAST elnevezésű szoftvere) segítenek abban, hogy felfedjük a vizsgált szekvenciánk és az NCBI adatbázisában tárolt szekvenciák vagy a saját kérdéses szekvenciáink közti hasonlóságokat.

Ha például kíváncsiak vagyunk arra, mennyiben egyezik meg egy általunk kiválasztott szekvenciarészlet a hemoglobin-B-t és hemoglobin-D-t kódoló génszakaszokon, a vizsgált szekvenciák 95%-os egyezését tapasztaljuk *(1. ábra)*.

```
Identities = 33/35 (95%), Gaps = 0/35 (0%)
 Strand=Plus/Plus

Szekvencia(1) (HBB)     3 GGCTGCCCTAACACCCCATGGGATGACACGGGATG  37
                            ||||  ||||||||||||||||||||  |||||||||
Szekvencia(2) (HBD)     3 GGCTCCCCTAACACCCCATGGGATGGCACGGGATG  37
```

*1. ábra.* Példa a páronkénti szekvenciaillesztésre (http://blast.ncbi.nlm.nih.gov/Blast.cgi)

***Evolúcióbiológia***

A bioinformatika segítségével ma már nemcsak fenotípusbeli jellemzők alapján, hanem a genomikai/proteomikai adatbázisokban tárolt információk elemzésével is vizsgálhatjuk az evolúciós folyamatokat.

A nukleotid- és fehérjeszekvenciák illesztésével ugyanis felfedhetők a különböző organizmusok közti evolúciós kapcsolatok. Eddigi tapasztalataink szerint az egymással közelebbi rokonságban álló fajok szekvenciái nagyobb hasonlóságot mutatnak, mint az evolúcióban korábban szétvált fajoké. Míg két ember genetikai állománya 99,9%-ban azonos, ez az érték ember és csimpánz közt 97-99%, ember és szarvasmarha között pedig mintegy 80%.

Ennek az az oka, hogy a génekben különböző mutációk halmozódnak fel, és minél több idő telik el a „közös” gén hordozása óta, annál inkább különbözni fog a két szekvencia. Ráadásul a hemoglobin esetében az aminosavak cseréjének mértéke nagyjából arányban is van az elválás óta eltelt

idővel (bár ez a folyamat nem tekinthető törvényszerűnek, hiszen az aminosavcserék sebessége általában sok paraméter függvénye, így nem állandó).

A többszörös szekvenciaillesztés (több mint 2 szekvencia illesztése) lehetőséget ad arra is, hogy a különböző szervezetekből származó, de azonos funkciójú fehérjék szekvenciáinak konzerválódott, funkcionálisan fontos és jellemző részleteit (pl. katalitikus helyeket, fehérje-fehérje kölcsönhatásban részt vevő motívumokat) felismerjük. Többek között ezeket a konzervált motívumokat/doméneket használják fel a kutatók a fajok közti evolúciós kapcsolatok rekonstruálásához, és ezek alapján képesek megállapítani a fajok evolúciós eltávolodásának hozzávetőleges idejét. Az igen összetett analízis eredményeit filogenetikai fán szokás ábrázolni.

Ezen túlmenően a konzervált domének felhasználhatók a fehérjék klasszifikációjának, azaz a rokon fehérjéket tartalmazó osztályok kialakításának folyamatában is.

Például a hemoglobin-B molekula fehérjeszekvenciájában található 147 aminosav hosszúságú szakasz 99-100%-os egyezést mutat több száz különböző eukariótaszekvenciával. Ez a szakasz megtalálható a konzervált domének adatbázisában is (CDD, http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). Azok a fehérjék, amelyek ezt a domént tartalmazzák, a globinszupercsalád tagjai.

***Fehérjeszerkezet-predikció***

A szekvenciaillesztésnél leírt módon, a fehérjeszerkezet ismeretének hiányában lehetőségünk van a 3D térszerkezet megjósolására, hasonló aminosavsorrendű és ismert szerkezetű fehérjék vagy fehérjealegységek (domének) felhasználásával. Számottevő szekvenciaazonosság hiányában a térszerkezet-predikciót fizikai elvek figyelembevételével végezzük el.

***Génexpresszió-vizsgálat***

A különböző kórképekben eltérő mértékben aktiválódó gének vizsgálatára alkalmasak a microarray- (chip-) módszerek, amelyekben a génekről átíródó RNS mennyiségét (expressziójának mértékét) hasonlítjuk össze a különböző (pl. egészséges kontroll) mintákban. Az expressziós eltérésekből következtethetünk az adott betegség kialakulásában szerepet játszó fontosabb génekre. Mivel egy-egy ilyen microarraylemezen több tízezer gént tudunk egyidejűleg vizsgálni, már egy mérés során akkora adathalmaz keletkezik, amelyet képtelenek lennénk informatikai támogatás nélkül feldolgozni és értelmezni.

A kapott eredményeinket feltölthetjük génexpressziós adatbázisokba, és mások eredményei között is keresgélhetünk. Ilyen adatbázis például a GEO (Gene Expression Omnibus): http://www.ncbi.nlm.nih.gov/geoprofiles

Mivel akár több ezer gén expresszálódik különböző mértékben egy-egy ilyen mérés során, érdemes a már említett hálózatokba illesztve vizsgálnunk az eredményeinket.

A különböző gének részvételét és kölcsönhatásaikat a metabolikus és jelátviteli útvonalakban, illetve molekuláris hálózatokban az IPA (Ingenuity Pathway Analisis) programmal tanulmányozhatunk (http://www.ingenuity.com/).

Az IPA támogatást nyújt egyrészt a gének egymáshoz való viszonyának értelmezésében, de segítségével – az új mérési eredményeink tükrében – bővíthetjük, módosíthatjuk a meglevő útvonalakat, folyamatosan bonyolítva ezzel a kutatók által feltárt összefüggéseket és hálózatokat.

Az expressziós microarrayvizsgálat példáján értelmezhető a bioinformatika alapvető szerepe az egyéb high throughput, azaz egy mérés során adatpontok ezreit generáló genetikai (pl. SNP, CNV), epigenetikai (pl. metilációs mintázat) és proteomikai vizsgálatoknál is.

## SZÓSZEDET

**Sanger-féle nukleotidszekvenálás:** egy DNS-szakasz nukleotid sorrendjének meghatározására használt módszer, amely során új DNS-szálat szintetizálunk a szekvenálandó szakaszból származó egyszálú templátot és szabad nukleotidokat felhasználva. Az új szál szintézise bármikor megállítható úgy, hogy nem a hagyományos bázisokat (4-féle dNTP-t), hanem azok didezoxiszármazékát (ddNTP) adják a rendszerhez. Radioaktív vagy fluoreszcens jelölést alkalmazva minden megállításnál leolvasható, hogy éppen melyik ddNTP épült be, ez utal a kiindulási DNS szál bázissorrendjére.

**Első genomprogramok:** a Bacteriophage MS2 RNS-genomjának megszekvenálását (1976), illetve a phi X 174 bakteriofág 5386 bázispár hosszúságú DNS-genomjának Sanger-módszerrel történt megszekvenálását (1977) értjük rajta.

**Genotipizálás:** a vizsgált organizmusra jellemző genotípus meghatározása. A használt módszertől függően jelentheti egyetlen nukleotid meghatározását (SNP-vizsgálat), ismétlődő elemek számának megállapítását, illetve az egyre terjedő teljesgenom-vizsgálatok esetében egyszerre több száz/több ezer SNP vagy hosszabb-rövidebb DNS-szakasz „leolvasását" is. A szekvenálás is a genotipizálás egyik robusztus módszere.

**Multifaktoriális betegség:** genetikai, epigenetikai és környezeti hatásra kialakuló kórkép. Összesített populációs gyakoriságuk magas, a betegségek változó súlyosságúak lehetnek. Öröklésük nem hasonlít a monogénes kórképekéhez. Idesorolható a magas vérnyomás, az elhízás, a cukorbetegség, a depresszió, a skizofrénia, az asztma stb.

**Epigenetikai módosulás:** olyan DNS-szerkezetváltozás, amely – szemben a klasszikus mutációval – nem jár a bázisszekvencia megváltozásával. Ismereteink szerint a leggyakoribb formája a DNS négy bázisa közül az egyiknek, a citozinnek a metilálódása. A metilálódás következtében az adott szakasz genetikailag inaktív lesz, vagyis a gén funkciója kiesik. A klasszikus mutációval ellentétben ez azonban reverzibilis folyamat, amely meghatározza a gének ki-be kapcsolását. A környezeti hatások, a táplálkozás, a magzati korban az anya dohányzása mind befolyásolják az epigenetikai mintázatot, ami a sejtosztódások során, akár generációkon keresztül is fennmaradhat.

**Proteomika:** a proteom, vagyis az élő szervezetben előforduló összes fehérje megismerésével foglalkozó tudományterület.

**Biobankok:** élőlényekből származó biológiai minták gyűjteményei. Típusát tekintve sokféle lehet, például: vér, tumoros szövetek, fehérje, DNS, RNS. A biobankokban felhalmozott minták tárolása és dokumentációja szigorú szabályokat követ, ennek következtében a minták felhasználhatók mind kutatási, mind pedig terápiás célra is (pl. vérátömlesztés).

**Homológia:** közös evolúciós eredetből származó nagyfokú szekvenciahasonlóság.

**Genomannotáció:** a szekvencia-adatbázisokban tárolt információ felruházása „értelemmel”, azaz a funkcionális szakaszok (gének, szabályozó elemek) meghatározása.

**Repetitív szekvencia:** hosszabb-rövidebb ismétlődő szekvenciák a genomban, amelyekben az ismétlődések száma diagnosztikus értékű lehet. Például a Huntington-kór esetében egy trinucleotid repeat (CAG) kóros felszaporodása figyelhető meg. Egészséges személyeknél a CAG-ismétlődésszáma 9–36 között váltakozik, a huntingtonos betegek esetében pedig ez 36–121 értékek közé esik, az adott gén funkcióváltozását okozva.

**High-throughput technikák:** „nagy áteresztőképességű” mérési eljárások, egy vizsgálat során párhuzamosan több száz/több ezer mérést végeznek el. Például egy ember akár egymillió SNP-jét is tudjuk egyszerre (egy array lemezen), rövid idő alatt vizsgálni.