Magyar Tudomány 181(2020)1, 116–120
DOI: 10.1556/2065.181.2020.1.13

Tudós fórum

BIOINFORMATIKA-OKTATÁS ÉLETTUDOMÁNYI HALLGATÓK RÉSZÉRE A GALAXY-PLATFORM SEGÍTSÉGÉVEL

TEACHING BIOINFORMATICS FOR STUDENTS OF LIFE SCIENCE ON THE GALAXY PLATFORM

Bálint Bálint László[1], Scholtz Beáta[2]

[1]PhD, MD, egyetemi adjunktus, Debreceni Egyetem Általános Orvostudományi Kar Biokémiai és Molekuláris Biológiai Intézet, Debrecen
lbalint@med.unideb.hu

[2]PhD, egyetemi docens, Debreceni Egyetem Általános Orvostudományi Kar Biokémiai és Molekuláris Biológiai Intézet, Debrecen

ÖSSZEFOGLALÁS

A genomikai adatgenerálás robbanásszerűen megnövelte a biológiai rendszerek jellemzésére rendelkezésre álló adatmennyiséget. Az óriási adatmennyiség feldolgozása lassúbb, mint az adatgenerálás sebessége, és a sebességet meghatározó lépés a rendelkezésre álló humán erőforrás. A bioinformatikai szakemberek döntően informatikai-matematikai alapképzettséggel vagy biológiai alapképzettséggel rendelkeznek, és a két területen használt nómenklatúrák, gondolkodási sémák akadályai a sikeres együttműködések kialakításának. Az elmúlt évtizedben alakult ki, és mára érett adatelemzési környezetté vált a Galaxy platform, amely sikeresen tudja ötvözni a parancssoros és grafikus felületen megvalósuló genomikai adatelemzést. A Galaxy platform bevezetése a bioinformatikai oktatásba sikeresen járulhat hozzá a bioinformatikai szakemberképzés sikeréhez, és ezáltal felgyorsíthatja ezen határterület fejlődését.

ABSTRACT

The boom in genomic data acquisition technologies increased dramatically the available datasets that can characterize biological systems. The speed of data processing is significantly behind the speed of data generation, and a major rate-limiting step is the availability of the human resource component. Experts in the bioinformatic analysis have a background in the mathematics-informatics field or in biological sciences. These two fields use different nomenclatures and different conceptual frameworks that can be real obstacles to successful cooperation. In the last decade the Galaxy Data Analysis Platform became a mature environment that is successfully integrating command-line programs with a graphical user interface based data analysis. Introducing the Galaxy Platform in the teaching of bioinformatics can improve the efficiency of teaching and therefore could speed up the overall development of the field.

Kulcsszavak: bioinformatika, nyílt forráskód, Galaxy, genomika, oktatás, képzés

Keywords: bioinformatics, open source, Galaxy, genomics, education, training

Az élettudományok területén az elmúlt évtized nagymértékű adatrobbanáshoz vezetett, és egyre szélesedik a szakadék a genomikai adatok előállítása és feldolgozása között. Az adatelőállítás költsége a Moore-törvényt (Gordon Moore, 1965) követve, sőt meghaladva csökken (November, 2018), de az adatfeldolgozás sebessége ezzel nehezen képes lépést tartani. Az akadémiai szféra informatikus szakemberek bevonásával nagyon korán létrehozta, és folyamatosan fejleszti az adatelemzéshez szükséges bioinformatikai és biostatisztikai eszközöket – ezek ugyan szabad felhasználásúak és a kutatói közösség számára hozzáférhetőek, de használatuk informatikai előképzettséget igényel. A tapasztalat azt mutatja, hogy az élettudományi kutatói közösség programozási készségeit időigényes és nehéz felépíteni, ezért az adatfeldolgozás limitáló tényezője mára már egyértelműen nem technikai, hanem humánerőforrás jellegű.

A bioinformatikai oktatás mint határterület jelentős kihívásokkal szembesül. A bioinformatikában dolgozó szakemberek vagy élettudományi háttérrel, vagy adatelemzési-matematikai-statisztikai-informatikai háttérrel rendelkeznek. A graduális képzési rendszer sajátosságai miatt a biológia-élettudományi alapképzettséggel rendelkező hallgatók és a matematikai-informatikai alapképzettséggel rendelkező hallgatók egyetemi tanulmányaik előtt más készségprofillal rendelkeznek, és ez a készségprofil az egyetemi képzési irányuknak megfelelően erősödik tovább. Ebből a különböző készségprofilból adódóan, a bioinformatikai oktatásban a matematikai-informatikai szemlélet elsajátítása az élettudományi alapokkal rendelkező hallgatók számára nagyon jelentős kihívást jelent. A fordított megközelítés is komoly nehézségekbe ütközik, nevezetesen amikor élettudományi szemléletet használva oktatnak matematikai-informatikai készségekkel rendelkezőket. A hétköznapi életben ezt nyelvi-kommunikációs falként élik meg azok a szakemberek, akik a két tudományterület képviselőjeként szakmai egyeztetési helyzetbe kerülnek. A nyelvi gát egyébként valóságosan létezik, hiszen azonos feladatokra a két területen más és más kifejezéseket használhatnak – jellemzően az informatikai szaknyelvben tágabban értelmezhető kifejezéseket, mint például az „adattisztítás", a genomikai elemzésben folyamatspecifikusan különböző nevekkel jelölhetnek (például: „variant calling", „peak-calling" stb.).

Az adatelemzési kihívások kezelésére több válasz született, mely válaszok eltérő filozófiák mentén különböző szoftverfejlesztésekhez vezettek.

Az élettudományi kutatói közösség adatelemzési igényeire az egyik választ az ipari fejlesztők által kínált, meglehetősen drága integrált adatelemző szoftverrendszerek jelentik, melyek használata programozói tudást nem igényel. Egy-egy ilyen szoftverrendszer jellemzően éves előfizetési díjjal érhető el, amely akár több ezer eurós kiadást is jelenthet. Globális szinten néhány száz előfizető biztosítja a folyamatos fejlesztéshez szükséges erőforrásokat.

Egészen más filozófiát képvisel a 2005-ben létrehozott Galaxy-projekt, melynek központi eleme egy ingyenesen használható, grafikus kezelői felületen ke-

resztül elérhető és elemzési funkciókkal rendelkező adatfeldolgozó rendszer, a Galaxy-platform (Giardine et al., 2005; Afgan et al., 2018). A grafikus felület miatt a Galaxy-platform használata nem igényel programozói tudást, így valósítja meg a projekt egyik fő célkitűzését, az általános hozzáférhetőséget. Ugyanakkor viszont a rendszer bioinformatikai eszközei, programjai a bioinformatikusok számára a központi tárhelyen keresztül szabadon elérhetők, ami biztosítja az elemzések átláthatóságát, és megkönnyíti a fejlesztéseket is. A Galaxy-platform jelenleg több mint nyolcvan ingyenes szerveren és ennél is több nem publikus szerveren működik, és több száz akadémiai kutató közösen fejleszti – ez a több lábon álló, sokszínű, eleven fejlesztői és oktatói közösség egyfajta biztosíték is a rendszer fennmaradására. A Galaxy-platform elérhetősége és sikeressége alapjaiban kérdőjelezi meg a bioinformatikai szoftverek üzleti modelljét.

A Galaxyt eredetileg a genomikai adatvizualizációs platformokkal párhuzamosan fejlesztették (UCSC Genome Browser, Ensembl), és korán összekapcsolták nagy genomikai projektekkel, mint például az ENCODE (Blankenberg et al., 2007). Mivel azonban a Galaxy egy olyan integráló rendszer, amelybe elvileg bármilyen adatfeldolgozó program beilleszthető, ezzel a háttérrel ma már szinte minden típusú *big data* analízis kivitelezhető. Molekuláris biológusok számára változatlanul a genomikai és transzkriptomikai elemzések a legizgalmasabbak – ideértve a genomannotációt, az összehasonlító genomikát és a metagenomikai elemzéseket, a mutációk és polimorfizmusok elemzését vagy az összehasonlító génexpressziós elemzéseket, az alternatív *splicing* és a ChIP-szekvenáláson alapuló génexpressziós szabályozás témaköreit. Ezeken túlmenően a Galaxy-projektben intenzíven fejlesztik a proteomikai és metabolomikai platformokat, illetve léteznek már ökológiai, kémiai és képelemző Galaxy-szerverek is (Batut et al., 2018). A legújabb fejlesztések pedig a gépi tanulási programokat (machine learning) integrálják a Galaxy-platformba. A rendszer nagyon rugalmasan kezeli az adat- és a munkafolyamat-megosztásokat, és lehetővé teszi a szabványos adatfeldolgozást, ami jelentősen elősegíti a tudományos eredmények egyértelmű kommunikálását és az elemzések reprodukálhatóságát.

A Galaxy-rendszer egyik erőssége, hogy interakciós felületet biztosít a matematikai-informatikai háttérrel rendelkezők és a biológiai háttérből jövő kutatók számára. A grafikus felhasználói felület mögött minden kód elérhető, és a bioinformatikai szoftverek a Galaxy-platformtól függetlenül is használhatóak. Ez a nyitottság lehetőséget ad az élettudományi területről érkezők számára továbblépni az R és UNIX szoftverek használatára. Az átmenet fokozatos lehet – a hallgatók/kutatók a genomszekvenálási technológiák ismeretében könnyen átlátják az adatelemzési folyamatok logikáját, és a Galaxy platformon keresztül megismerhetik a szoftverek lehetőségeit és korlátait, valamint kombinálhatóságukat különböző elemzési célok eléréséhez. Ez jelentősen megkönnyíti, hogy továbblépjenek a parancssoros elemzések elsajátítása irányába, vagy továbbra is a Galaxy-plat-

formon használhatják a beállított munkafolyamatokat. Bizonyos Galaxy-platformon megismert szoftvereket teljes eszköztárukkal és maximális flexibilitásukkal a parancssoros alkalmazásban lehet használni.

A Galaxy-rendszeren alapuló analízisek megtanulására több lehetőség is van:

1. rövid, 1-2 napos *workshop*ok, melyeket különböző egyetemek bioinformatikai kutatócsoportjai szerveznek, célzottan egy bizonyos típusú adatelemzés oktatására;
2. az évente megtartott Galaxy Community Conference, melyet felváltva Európában, illetve az Amerikai Egyesült Államokban szerveznek, és – egyéb tudományos programok mellett – a képzések során többféle adat elemzését oktatják, kezdő és haladó szinten;
3. önképzés, a Galaxy projekt saját, színvonalas oktatási segédanyagait (Galaxy Training, URL1), és egyéb, az interneten elérhető publikációkat és oktatóvideókat használva.

A legnagyobb európai Galaxy bioinformatikai szerver a UseGalaxy.eu Freiburgban érhető el (University of Freiburg, Németország, Freiburg Galaxy Team), és az ELIXIR-EUROPE hálózat keretében működik. Itt új szolgáltatásként (egyelőre ingyenesen) biztosított a „Training Infrastructre As a Service", a Galaxy-képzések számára elkülönített szerverhozzáférés. 2019-ben a központi Galaxy-szervert egy európai hálózatba kapcsolták, mely lehetővé teszi, hogy a forráselosztás révén a felhasználók kihasználása gördülékenyebb legyen. A hálózatba kapcsolt szerverek különböző helyszíneken vannak, de a programcsomagok a Github-platformon keresztül szinkronizálásra kerülnek. Friss fejlemény, hogy kidolgozásra került egy olyan munkafelület is, melybe a parancsoros programok kódjai tetszés szerint importálhatók és összekapcsolhatók a grafikus felhasználói felülettel.

Magyarországon a Galaxy-alapú bioinformatikai képzéseket dr. Scholtz Beáta vezette be 2016-ban a molekuláris biológus MSc-képzés keretében, másodéves hallgatók számára. Ezen túlmenően 2018-ban a Debreceni Egyetemen két magyar nyelvű Galaxy-képzést tartottunk, kutatók és PhD-hallgatók számára, illetve tartottunk egy ELIXIR-kompatibilis, angol nyelvű Galaxy-képzést is, ugyancsak dr. Scholtz Beáta vezetésével. A képzések fő fókusza eddig az újgenerációs RNS szekvenálási adatok elemzése volt, és javarészt élettudományi kutatásokban aktív PhD-hallgatók, illetve néhány szenior munkacsoport-vezető és ipari kutatók részvételével zajlottak.

A képzés megszervezését az EFOP-3.6.1 kódszámú „Intelligens szakosodást szolgáló intézményi fejlesztések" című pályázati konstrukció keretében megvalósuló „Debrecen Venture Catapult Program" tette lehetővé. A képzés által az ELIXIR kutatói hálózathoz való kapcsolódás és a kutatói utánpótlás fejlesztése terén sikerült a pályázati tevékenységekhez hozzájárulnunk. A képzésekhez részben a Debreceni Egyetemen működő, limitált kapacitású Galaxy-szervert

vettük igénybe – nagy segítséget jelentett azonban, amikor a helyi szerver technikai problémái miatt a képzés megakadt, rövid idő alatt és akadálymentesen a képzés technikai hátterét át tudtuk váltani a freiburgi Galaxy-szerverre (URL2). A program keretében elkészült egy magyar és egy angol nyelvű Galaxy gyakorlati jegyzet is, amely lefedi az RNS szekvenálási lépések elméletét, gyakorlatát és a legfontosabb használandó programcsomagok ismertetését is.

Bízunk benne, hogy a hazai élettudományi közösség számára egyre szélesebb körben lesz elérhető és megismerhető a Galaxy bioinformatikai rendszer, melynek használata által jelentősen javulhat a bioinformatikai oktatás hatékonysága és a big data élettudományi hasznosításának üteme.

KÖSZÖNETNYILVÁNÍTÁS

A publikáció elkészítését és a képzések megvalósítását az EFOP-3.6.1-16-2016-00022 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg. Bálint B. L. az ELIXIR-Magyarország (http://elixir-hungary.org/) oktatási és képzési koordinátora.

IRODALOM

Afgan, E. et al. (2018): The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update. *Nucleic Acids Research*. Narnia, 46, W1, W537–W544. DOI: 10.1093/nar/gky379, https://academic.oup.com/nar/article/46/W1/W537/5001157

Batut, B. et al. (2018): Community-Driven Data Analysis Training for Biology. *Cell Systems,* 6, 6, 752–758.e1. DOI: 10.1016/j.cels.2018.05.012, https://bit.ly/2r85MH5

Blankenberg, D. et al. (2007): A Framework for Collaborative Analysis of ENCODE Data: Making Large-scale Analyses Biologist-friendly. *Genome Research,* 17, 6, 960–964. DOI: 10.1101/gr.5578007, https://genome.cshlp.org/content/17/6/960.long

Giardine, B. et al. (2005): Galaxy: A Platform for Interactive Large-scale Genome Analysis. *Genome Research,* 15, 10, 1451–1455. DOI: 10.1101/gr.4086505, https://genome.cshlp.org/content/15/10/1451.long

Moore, G. E. (1965, 2006): Cramming More Components onto Integrated Circuits. Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff., *IEEE Solid-State Circuits Society Newsletter,* 11, 3, 33–35. DOI: 10.1109/N-SSC.2006.4785860, https://newsroom.intel.com/wp-content/uploads/sites/11/2018/05/moores-law-electronics.pdf

November, J. (2018): More than Moore's Mores: Computers, Genomics, and the Embrace of Innovation. *Journal of the History of Biology,* 51, 4, 807–840. DOI: 10.1007/s10739-018-9539-6, https://link.springer.com/content/pdf/10.1007/s10739-018-9539-6.pdf

URL1: https://training.galaxyproject.org
URL2: https://usegalaxy.eu