UDC: 004.451.9 Info M str. 26-34

БИО-ЛИНУКС У СВЈЕТЛУ САВРЕМЕНИХ БИОИНФОРМАТИЧКИХ ТЕНДЕНЦИЈА
BIO-LINUX IN THE LIGHT OF MODERN BIOINFORMATIC TENDENCIES

Сретенка Видић, Дејан Кременовић, Димитрије Д. Чвокић

РЕЗИМЕ: Представљена је дистрибуција Линукса звана Био-Линукс, као општа биоинформатичка платформа превасходно намијењена биолозима истраживачима и биоинформатичарима. На почетку је дат краћи приказ биоинформатике као научне дисциплине, а потом су предочени разлози који су утицали на настанак и развој оперативног система Био-Линукс, као и приказ неких од његових најбитнијих карактеристика. Извршено је поређење са осталим, у литератури познатим, био-оријентисаним дистрибуцијама, али и са двије тренутно најпопуларније научно-оријентисане дистрибуције: Scientific Linux и Fedora Scientific. Пред крај је дат осврт на Био-Линукс и из дидактичког угла. Циљ рада је да упозна ширу стручну заједницу са потенцијалом и могућностима Био-Линукса, те да укаже на неке аспекте који би можда били од користи за развој биоинформатике код нас.

КЉУЧНЕ РЕЧИ: Линукс, оперативни системи, биоинформатика, биолошки софтвер

ABSTRACT: We give an overview of Linux distribution called Bio-Linux, as a general bioinformatics platform meant for use by biologists and bioinformatics scientists. In the beginning, we give a short overview of bioinformatics, its history and development. After that, we present the reasons that have contributed to the emergence and development of the Bio-Linux operating system, as well as some of its main features. We compared this one with others bio-oriented distributions that are known in the literature and with the two most popular scientific distributions today: Scientific Linux and Fedora Scientific. In the end, we emphasised the didactical role of Bio-Linux. The goal of this paper is to inform the broader professional community in the western Balkans with the potentials of Bio-Linux OS and to point out some aspects that might be useful for the bioinformatics development in the region.

KEY WORDS: Linux, operating systems, bioinformatics, biological software

УВОД

Рачунарство, као дисциплина, увелико је постало неотуђив дио савремене цивилизације. Нове технологије уводе се непрестано и при томе застарјевају готово чим се појаве (1). Већ 45 година уназад се веома активно проучавају принципи дјеловања биолошких материја на изоловане органе код људи и животиња са посебним нагласком на информационо стање ћелија. Термин „биоинформатика" је први пут поменут 1970. г. у раду Хеспера Хогевега под називом „Bioinformatica: een werkconcept" (2). Означавао је било какву примјену информационих технологија за потребе биологије. Током 80-их, истраживања постају превасходно усмјерена на ћелију као функционалну јединицу и такозване међућелијске комуникације. Упоредо са истраживањима у генетици, чији је циљ посвећен проучавању механизама насљедних информација, дошло је до издвајања биоинформатике као новог истраживачког правца, а сам термин почиње да се односи углавном на алгоритме за ДНК секвенцијалну анализу (3). Непосредни значај биоинформатике за човјека, природу и квалитет живота је велики, пошто резултати биоинформатичких истраживања доприносе побољшаној дијагностици болести, откривању генетске предиспозиције болести, дефинисању генске терапије, раном откривању и третману патогенеза, идентификовању ДНК, и сличном (за више информација погледати (4) и (5)).

Анализа поменутих ДНК секвенци врши се из читавог низа подобласти примјеном великог броја различитих метода и алгоритама. Основна средства биоинформатике су статистичке анализе расподјеле слова биграма, триграма, заступљености кодона, ентропије, анализа зависности, и томе слично. До данас је развијен велики број различитих информационих система, база података, али и других биоинформатичких ресурса за њихову анализу. Према (6), главни биолошки проблеми на које се фокусира биоинформатика су: молекуларни и ћелијски механизами, развој комплексних организама, и еволуција. С тим у вези, одговарајући (био)информатички проблеми постају: састављање геномских секвенци, анотација, поравнање геномских секвенци, одређивање 3Д структуре протеина, реконструкција филогенетских веза међу врстама, прорачуни утицаја лијека или токсина на организам, те симулације биолошких процеса приликом испитивања одговарајућих математичких модела.

С друге стране, свака биоинформатичка анализа захтијева и одговарајућу рачунарску платформу. Стога се поставља питање шта то треба да карактерише идеалну рачунарску платформу кад је ријеч о биоинформатици? Кажу да је „љепота у очима посматрача", тако да ни овај случај не би требао бити изузетак. Генерално, ствари на које се обраћа пажња су: брзина, стабилност, сигурност, удобност, могућност умрежавања, ниска цијена, али и бројност савремених програма за научна израчунавања и обраду података који се могу извршавати на самој платформи. Иначе, због све веће примјене тзв. секвенцирања новог покољења (СНП) (енг. new generation sequencing (NGS)) и жеље за што бржим добијањем повратних резултата, постало је практично неизводљиво геномску анализу предавати у руке искључиво специјалистима на том пољу. Умјесто тога, сами истраживачи преузимају на себе бар је-

дан дио посла. Но, за тако нешто морају бити испуњена бар сљедећа два услова:

- омогућен приступ и рад са одговарајућим софтвером
- истраживачи морају усвојити одређене вјештине како би били у стању да се изборе са огромним количинама СНП-података.

Штавише, уобичајени начини обраде података врло брзо застарјевају у ери вртоглавог развоја нових технологија за секвенцирање, нових лабораторијских протокола, нових питања која се рађају, нових референтних база података, као и раста обима података. Конкретно, секвенцирање генома од 1000$ може да захтијева анализу од 100 000$ (7). Даље, у биологији дошло до праве експлозије различитих типова података и формата датотека, а управо је један од задатака рачунарских и информационих система да такве, сирове, необрађене податке похране и преведу у погодне математичко-комбинаторне структуре које би касније омогућиле колико-толико дјелотворну обраду и њихову интерпретацију. Стога, поменуте карактеристике нису тек некаква хировита жеља истраживача у области тзв. животних наука, већ све више и више постају од суштинске важности за реализацију самих истраживања.

Испуњењу захтјева за идеалном софтверском платформом најближи су бесплатни софтверски системи са јавно доступним кодом (енг. Free and open source software (FOSS)). Можемо рећи да су највећи биоинформатички пројекти на самим својим почецима нашли уточиште у FOSS софтверу (8). На примјер, према Линколну Штајну, истраживачу из Колд Спринг Харбор Лабораторије, програмски језик Перл је био просто спасоносан за пројекат секвенцирања људског генома, што је и образложио у свом чланку из 1996. г. насловљеном „How Perl saved the Human Genome Project" (9). Други примјер великог пројекта који се увелико ослањао на FOSS софтвер је секвенцирање SARS вируса. У чланку Мартина Крзивинског „Sequencing the SARS Virus", објављеног 2003. године у часопису LINUX Journal, описан је цјелокупан поступак секвенцирања који се вршио на RedHat Linux дистрибуцији, уз коришћење MySQL-a, Apache веб-сервера, и добро познатих биоинформатичких апликација BLAST, Phred, Phrap и Consed (10). Трећи примјер би могао да буде Пројекат европског отвореног софтверског пакета за молекуларну биологију – EMBOSS пројекат (енг. European Molecular Biology Open Software Suite). Сам пакет се састоји од око 300 апликација и подржава различита графичка и апликацијска програмска прочеља (енг. Graphical User Interface – GUI и Application Programming Interface – API). Популарност FOSS софтвера је непосредно узроковала и формирање Фондације за отворену биоинформатику (енг. Open Bioinformatics Foundation – OBF), као непрофитне организације усредсређене на подршку биоинформатичком софтверу са јавно доступним програмским кодом. У складу са тим, чувени часопис Bioinformatics, основан још 1985. г., је у јулу 2005. г. на свом веб-сајту убацио допунску инструкцију ауторима: „Софтвер и подаци морају бити јавно доступни некомерцијалним корисницима. Доступност мора бити изричито наведена у раду. Аутори морају, такође, да обезбиједе да је њихов софтвер јавно доступан током цијеле двије године од објављивања рада. Веб-сервиси не би требали да захтијевају регистрацију. Додатне и допунске податке на интернету може да објави искључиво часопис, или пак аутор на својој личној веб-страници.".

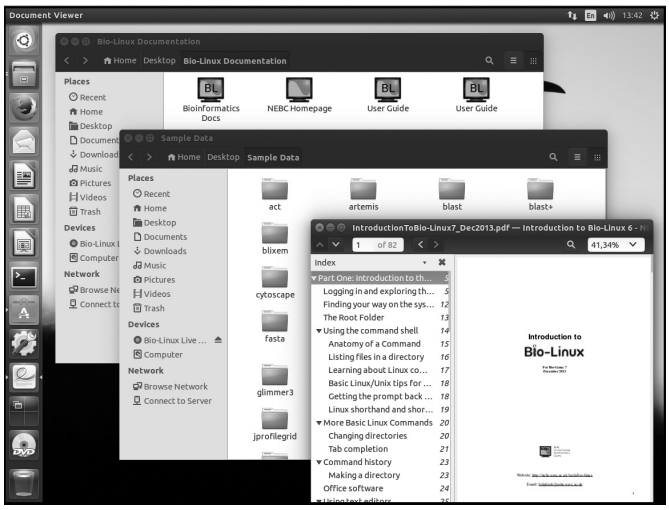
Иначе, кад је ријеч о FOSS софтверу, сасвим сигурно централно мјесто заузима Линукс (11). Због своје брзине, стабилности, сигурности, цијене (практично је бесплатан) и великог броја дистрибуција, према (12) Линукс је постао платформа која се користи на највећем броју машина врло разноликог хардвера: мобилни телефони, таблет рачунари, суперрачунари, мрежни сервери, и слично. За биоинформатичаре су Линуксови оперативни системи битни из простог разлога што је већина висококвалитетног биоинформатичког софтвера развијена управо за POSIX-системе. Штавише, под Линуксовом BASH љуском, отварање SAM/FASTQ датотека, које могу да заузимају и по 4GiB, и на примјер формирање нове састављене само од Chromosome 1 очитавања, може се извести помоћу једног grep-исказа, што је скоро незамисливо урадити под Виндоузом. То је довело до тога да се већ дужи низ година развијају различите дистрибуције које су превасходно биоинформатички оријентисане (8) (13) (14) (15). Једна од њих је и Био-Линукс (енг. Bio-Linux), чији је развојни пут започет у Великој Британији још 2002. г. у тадашњем NERC-центру за биоинфороматику и животну околину (енг. NERC Environmental and Bioinformatics Centre (NEBC)) (11). Иначе, честа употреба термина „Био" и „Линукс" у називима, изродила је и кровни термин БиоЛинукс (енг. BioLinux), који означава све пројекте који имају за циљ олакшани приступ биоинформатичком софтверу на Линуксу (11). У овом раду, поред претходно казаних неколико ријечи о бионформатици, биће представљен оперативни систем Био-Линукс, разлози због којих је настао, који су циљеви читавог пројекта, шта га то разликује од других оперативних система из Линуксове фамилије, и каква је његова улога као помоћног наставног средства у савременом биоинформатичком образовању.

БИО-ЛИНУКС КАО ОПШТА БИОИНФОРМАТИЧКА ПЛАТФОРМА

Већ у старту Био-Линукс се сврстао у ред бесплатног софтвера са јавно доступним кодом. Дозвола под којом се дистрибуира је Simplified BSD Licence (16). Убунту-база му обезбјеђује брзину, стабилност, могућност умрежавања, и смањену опасност од малициозног софтвера. Што се тиче ажурирања, популарност Убунтуа (16) му даје сигурност и на том пољу, што је грубо говорећи и својеврсан залог будућности. Посљедња верзија – Био-Линукс 8.0 – заснована је на Убунту 12.04 који спада у верзије са тзв. дугоркочном подршком (енг. Long Term Support (LST)). Иначе, униформност је једна од најбитнијих карактеристика за успјеш-

ну научно-истраживачку сарадњу, јер омогућује лакшу размјену података, промовише ширење најбоље праксе, и олакшава пружање централизоване и економичне подршке. Кад је ријеч о Био-Линуксу, техничка подршка је обезбијеђена путем електронског шалтера и радионица које организује NERC Environmental 'Omics Synthesis Centre (бивши NEBC). Велики број корисника, веома разноликог интересовања, представља непроцјењив ресурс за развој било каквог софтвера, па самим тим и за Био-Линукс. Програмерима је дозвољена програмска измјена карактеристика самог оперативног система и интеграција са новим компонентама и софтвером. Додаци (енг. plugins) за Убунту дистрибуцију се могу уградити у Био-Линукс без икаквих посебних модификација. Такође, пројекат укључује и формирање софтверских репозиторијума умјесто комплетних система, пошто је линукс-заједница доста поједноставила дистрибуцију софтвера преко „пакета", при чему је на тај начин сам корисник поштеђен њихове инсталације. Развој програма и њихово смјештање у репозиторијуме омогућује им да буду доступни свим корисницима, што у крајњој линији кроз употребу и критику доводи до сталног унапрјеђења. У пакету, заједно са Био-Линуксом долази преко 500 различитих биоинформатичких програма (17), а то га практично и чини једном биоинформатичком платформом.

Што се тиче ажурирања, као и већина дистрибуција ни Био-Линукс се не ослања на тзв. непрекидну доставу (енг. rolling release или continuous delivery) ажурираног софтвера, већ на тзв. развојне циклусе са унапријед дефинисаним роковима (18). Конкретно, изворни код овог биоинформатичког софтвера за Дебијан-Мед (енг. Debian-Med) компилује и пакује тим Дебијан-Меда. Послије тестирања на Дебијану (енг. Debian) ради се на одговарајућој верзији за Убунту, на основу које Био-Линуксов тим развија конкретну верзију Био-Линукса. Проблем је што се све одвија у циклусима од по двије године. Иако са становишта софтверског инжењерства доста повољнија опција, овакав приступ може довести до тога да на посљедњим верзијама Био-Линукса не буду инсталиране најновије верзије биоинформатичких програма. Да би се то превазишло, одлучено је да се формирају сопствени репозиторијуми за уградњу допунског софтвера, као и његово ажурирање. Штавише, нове верзије софтвера, које припреми Био-Линуксов тим, шаљу се назад Убунтуовим и Дебијановим тимовима на провјеру и допунску обраду. Постоји и други начин звани ретротрансфер (енг. backporting) који се односи на ручно преузимање пакета са најновије верзије оперативног система, преправку помоћу Убунтуових развојних оруђа (енг. Ubuntu-dev-tools), тестирање на старијим верзијама, отпремање на званичне Убунтуове сервере и писање извјештаја (за све мора да постоји одобрење главних пројектаната Убунутуа) (18).



Слика 1 *Радна површ Био-Линукса 8.0, прозори директоријума Bio-Linux Documentation и Sample Data, уз отворен документ Introduction to Bio-Linux.*

Када је ријеч о удобности, функционалност и изглед радне површи Био-Линукс дугује већ познатом Јунити (енг. Unity) графичком прочељу. Постоји засебан изборник за биоинформатички софтвер, те посебни директоријуми за документацију и тестне податке. На слици 1 дат је приказ графичког прочеља Био-Линукса 8.0 са отвореним прозором директоријума Bio-Linux Documentation и Sample Data.

И с концептуалне стране гледано корисницима је олакшан рад са софтвером, било да су се раније сусретали са Линуксом или пак нису. Инсталација различитог научног софтвера може да доведе до различитих врста колизије. У таквим ситуацијама је потребно веће техничко знање о самом софтверу, оперативном систему, процесу инсталације, и наравно вријеме за потребна подешавања. Био-Линукс, као оперативни систем и као колекција великог броја биоинформатичког софтвера, рјешава дати проблем нудећи све у пакету – истестирано, подешено, и синхронизовано. У табели 1 су приказани главни захтјеви који су дефинисали развој Био-Линукса, као и учињени кораци на том путу (11).

Табела 1 *Приказ захтјева који су дефинисали развој Био-Линукса.*

Кораци	Кориснички захтјеви NERC заједнице	NEBC Био-Линукс као рјешење
Дефинисање захтјева корисника	Довољно моћна и јефтина платформа кад је ријеч о истраживањима у области животне околине.	Покренут пројекат развоја оперативног система који ће да инкорпорира у себе широк спектар биоинформатичког софтвера.

Избор технологије	Јединствена рачунарска радна платформа која треба да олакша обраду, циркулисање, и прикупљање података, уз поједностављен и процес инсталације.	Дистрибуирање оперативног система заснованог на Линуксу, користећи се SystemImager-ом.
Прилагођавања	Сигурност и интегрисани backup-систем.	Специјални заштитни зид са помоћним диском за backup-е.
Додавање софтвера	Веома широка палета некомерцијалног биоинфор-матичког софтвера.	Одабрати више од 60 кључних биоинформатичких пакета.
Направити документацију	Свеобухватна и једноставна за коришћење документација.	Доступност документације путем интернета.
Објављивање тестног рјешења	Добро истестиран систем.	Започети коришћење у одабраној групи тестера.
Коначно презентовање система	Једноставна инсталација на захтјев.	Квалитетно структуиран и истестиран софтвер (систем) који представља основу за даљи развој.
Обука и сервис	Омогућити упознавање са Линуксом и биоинформатичким софтвером.	Организација курсева, развој веб-сајтова преко којих ће бити обезбијеђена стручну помоћ.
Одржавати као софтвер који развија заједница	Обезбиједити добру комуникацију са програмерима и корисницима.	Мијењати систем на основу повратних информација.

ПОРЕЂЕЊЕ БИО-ЛИНУКСА СА ОСТАЛИМ БИО- И НАУЧНО-ОРИЈЕНТИСАНИМ ДИСТРИБУЦИЈАМА

Постоји не тако мали број Линуксових дистрибуција које су измијењене и прилагођене потребама животних наука и биоинформатике (15) (13) (14) . Њихова бројност говори да одавно постоји жеља за униформном биоинформатичком платформом. Многе од њих представљају сад већ мртве пројекте, но обично се искуство стечено на њиховом развијању преносило на развој других. У табели 2 дат је преглед општих системских карактеристика различитих био-оријентисаних дистрибуција Линукса: 32/64-битна архитектура, минимална захтјевана количина RAM-а, да ли су бесплатне, јавна доступност кода, и какав им је статус. Критеријум за претпосљедњу карактеристику је била доступност пројекта путем неког од веб-базираних сервиса за похрану и верзирање (нпр. GitHub, SourceForge, Assembla), преко званичног веб-сајта дистрибуције, или пак ако је могуће прикључење самом пројекту. У противном, сматрали смо да код одговарајуће дистрибуције није јавно доступан. Наравно, тешко је очекивати доступност за деактивирани пројекат, без обзира каква је начелно била ситуација, јер независан тим програмера није у могућности да самостално покрене развој. Информацију о статусу смо базирали на подацима које смо могли добити путем званичних веб-сајтова и DistroWatch.com-а.

Табела 2 *Упоредни приказ општих системских карактеристика различитих био-орјентисаних дистрибуција Линукса*

	x86	64-битна архитектура	RAM	Бесплатан	Јавно доступан код	Активна дистрибуција
Био-Линукс	Да	Да	512 MiB	Да	Да	Да
Vlinux	Да	Не	512 MiB	Да	Не	?
Vigyaan	Да	Не	256 MiB	Да	Не	?
BioKnoppix	Да	Не	Нема података	Да	Не	Не
Quantian	Да	Не	512 MiB	Да	Не	Не
DNALinux	Да	Не	1GiB	Да	Не	Не
BioPuppy	Да	Не	256 MiB	Да	Не	Не
Bioconductor-Buntu	Да	Не	512 MiB	Да	Не	Не
PhyLIS	Да	Да	512 MiB	Да	Не	Не
Lxtoo	Да	Да	512 MiB	Да	Не	Не
Baari	Да	Не	Нема података	Да	Не	Не
BioSLAX	Да	Не	256 MiB	Да	Не	Не
BioBrew	Да	Не	512 MiB	Да	Не	Не
BioLinuxBR	Да	Не	Нема података	Да	Не	Не
BioLand	Да	Не	Нема података	Да	Не	Не
OpenDiscovery	Да	Не	512 MiB	Да	Не	Не

Можемо примијетити да је развој и одржавање већине дистрибуција деактивирано. Штавише, мало која од њих је подржавала рад на савременијим 64-битним машинама. За двије дистрибуције нема званичне потврде о њиховом статусу с обзиром на информације доступне путем одговарајућих веб-сајтова: VLinux и Vigyaan. Пошто нема смисла разматрати деактивиране пројекте, искључићемо их из даљег разматрања, али ћемо зато поређење проширити са двије опште научно-оријентисане дистрибуције (табела 3): Scientific Linux 7 и FedoraScientific 24.

Табела 3 *Приказ општих системских карактеристика за Scientific Linux и Fedora Scientific*

	x86	64-битна архитектура	RAM	Бесплатан	Јавно доступан код	Активна дистрибуција
Scientific Linux	Старије верзије	Да	1 GiB	Да	Да	Да
Fedora Scientific	Да	Да	1 GiB	Да	Да	Да

У посљедње вријеме виртуелизација узима све више маха, па је постојање одговарајућих .iso и .ova формата неминовност. Такође, сматрамо да је и важно да постоји могућност преузимања слике оперативног система и путем P2P (енг. peer-to-peer) технологије (нпр. BitTorrent, Gnutella, Freenet). Поред тога што је битно да буде доступна цјелокупна инсталациона верзија дистрибуције, пожељна је могућност коришћења у виду бутабилног флеш-драјва/DVD-а (тзв. готова варијанта). Приказ поменутих карактеристика дат је у табелама 4 и 5.

Табела 4 *Доступност дистрибуција.*

	.iso	.ova	P2P	Готова варијанта	Инстал. Верзија
Био-Линукс	Да	Да	Да	Да	Да
Vlinux	Да	Не	Не	Да	Не
Vigyaan	Да	Не	Не	Да	Не
Scientific Linux	Да	Да	Не	Да	Да
Fedora Scientific	Да	Не	Да	Да	Да

Табела 5 *Опште карактеристике одржавања система.*

	Посљедње издање	LTS	Допунски биоинформатички репозиторијум	Могућност ретротрансфера
Био-Линукс	2014	Да	Да	Да
Vlinux	2011	Не	Не	Не
Vigyaan	2005	Не	Не	Не
Scientific Linux	**2016**	Да	Не	Не
Fedora Scientific	**2016**	Не	Не	Не

Пошто се може примијетити да су VLinux и Vigyaan искључиво готове варијанте (енг. live versionc), а уз то и поприлично застарјеле, искључићемо их из даљег разматрања. У наредној табели имамо поређење које се у суштини односи на удобност при упознавању са биоинформатичким софтвером. Што се тиче ажурирања, можемо примијетити да вођство имају ScientificLinux и Fedora Scientific. За другопоменуту дистрибуцију је тако нешто и очекивано, с обзиром да је развојни циклус свега шест мјесеци. С друге стране, поменуте научно-оријентисане дистрибуције немају могућност ретротрансфера, нити за њих постоје допунски биоинформатички репозиторијуми, што је одлика Био-Линукса. Упроштено гледано, иако је посљедња верзија Био-Линукса угледала свјетло дана 2014. г., због дугорочне подршке, допунских репозиторијума и ретротрансфера спада међу благовремено ажуриране системе.

Табела 6 *Посебан изборник за биоинформатику, тестни подаци, подршка, и документација.*

	Изборник	Програми	Тестни подаци	Техничка подршка	Документација на веб-сајту	Документацијски директоријум
Био-Линукс	Да	~**500**	Да	Да	Да	Да
Scientific Linux	Не	0	Не	Да	Не	Не
Fedora Scientific	Не	0	Не	Да	Не	Не

Као што видимо из табеле 6, Scientific Linux и Fedora Scientific, иако врхунске научно-оријентисане дистрибуције, не садрже тестне податке нити пратећу документацију за биоинформатички софтвер.

Пред крај, ред је и да се види каква је ситуације кад је ријеч о конкретном бионформатичком софтверу. Наравно, тешко да бисмо били у могућности да на мало простора упоредимо шта све јесте и није могуће инсталирати на ове три дистрибуције. Ипак, у табели 7 приказали смо стање за један мали дио дате фамилије софтвера. Ријеч је о доступности неких од најпознатијих биоинформатичких пакета, како је то презентовано својевремено у раду (14). Напомињемо да су у том раду упоређиване друге дистрибуције, које данас већ можемо сматрати застарјелим (табела 2).

Табела 7 *Доступност неких од најпознатијих биоинформатичких пакета.*

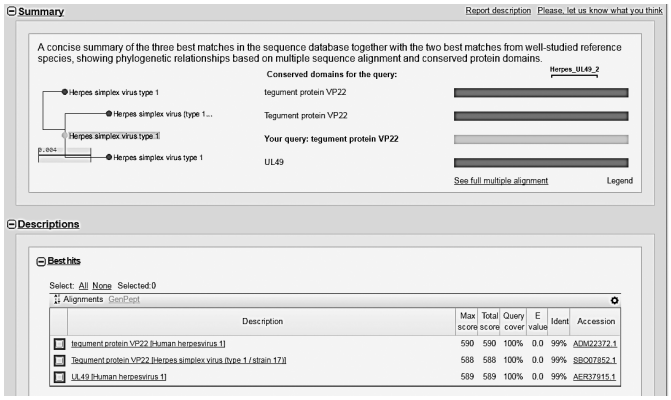
	EMBOSS	BLAST	Gromacs	Clustal Omega
Био-Линукс	Да	Да	Опционо	Опционо
Scientific Linux	Релативно компликована инсталација	Опционо	Релативно компликована инсталација	Опционо
Fedora Scientific	Опционо	Опционо	Релативно компликована инсталација	Опционо

Као што се може примијетити, Scientific Linux и Fedora Scientific представљају јако добре и погодне платформе и за биоинформатику, али ипак заостају када је ријеч о удобности, количини најновијег биоинформатичког софтвера, документацији, доступности тестних података, и синхронизацији инсталираног софтвера. Касније, у раду, биће ријечи и о неким другим карактеристикама које издвајају Био-Линукс у односу на ове двије, иначе, јако добре дистрибуције.

ПРИМЈЕР КОРИШЋЕЊА БИОИНФОРМАТИЧКОГ СОФТВЕРА НА БИО-ЛИНУКСУ

Релативно компликована инсталација EMBOSS-а може се сматрати значајнијим недостатком за неки озбиљнији

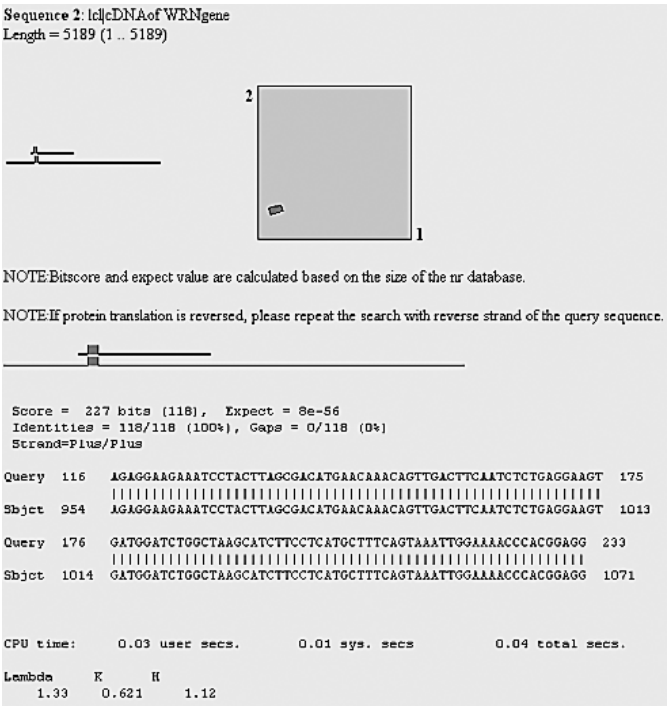
рад, кад је ријеч о Scientific Linux-у и Fedora Scientific-у. Но, изузимајући такве случајеве, сврсисходно би било упоредити поменуте дистрибуције и кад је ријеч о софтверу који се лако може инсталирати на њих. Примјер који слиједи се односи на употребу BLAST софтвера са којим су се већ сусретали аутори овог рада. BLAST представља оруђе за поређење примарних структура, било аминокиселинских секвенци протеина, или пак нуклеотида као примарне структуре одређене молекуле ДНК. На примјер, након открића претходно непознатог гена код миша, научници би могли коришћењем BLAST-а претражити постоји ли и у људском геному сличан ген (19). Иако на тржишту постоји много софтвера за претраживање и рад са базама података, ипак веб-ресурси доступни за BLAST спадају у најразвијеније, најстабилније, и имају приступ највећим базама података. Једна од предности BLAST програма јесте то што постоји организованост према типу секвенци које се претражују. На примјер, blastp програм упоређује протеинске секвенце са протеинским базама података, а blastn програм, с обзиром на проблем, враћа најсличније ДНК секвенце из ДНК базе података.



Слика 2 *Коришћење BLAST-а преко графичког прочеља.*

На слици 2 је дат примјер поређења насумично одабране протеинске секвенце са базом протеинских секвенци. Као што се може примијетити, дата секвенца се у потпуности поклапа са три протеина која су заступљена у људском вирусу херпеса. За мале количине података, гдје се желе претражити доступне базе података, веб-сервис је свакако добра опција. С друге стране, рад преко командне линије омогућује додатно подешавање претраге, обраду веће количине података, аутоматизацију, а и могућност приступа само жељеним информацијама (филтрирање).

На слици 3 је приказан резултат поређењеа двије секвенце – једне геномске, која представља дио из GenBank HTG записа, а који садржи и дио гена Вернеровог синдрома – ген WRN (eng. Werner's syndrome, WRN), и mRNK (cDNK) секвенци. WRN ген садржи 35 егзона. График BLAST-а приказује мапирање егзона у cDNK координатном систему. Нуклеотидне секвенце које се испитују, то јесте секвенце чија се сличност тражи, неопходно је „налијепити" на одређено поље. BLAST има за задатак да пронађе који је егзон, ако уопште постоји, садржан у HTG секвенци, када се већ поменута пореди са WRN геном cDNK секвенце. Из приложене слике 3 се јасно види да је нуклеотидни низ од 116-233 исти као и нуклеотидни низ од 954-1071 cDNK секвенце и тиме је утврђено да cDNK координате одговарају егзону број 8.



Слика 3 *Резултат поређења двије нуклеотидне секвенце у BLAST-у.*

Овај примјер се успјешно могао извршити на све три дистрибуције, али су Scientific Linux и Fedora Scientific захтијевале инсталацију BLAST-а и допунску конфигурацију, тј. подешавање такозваних „монтажних" промјењивих PATH (системска) и BLASTDB (софтверска), преузимање одговарајућих база података, валидацију, и томе слично. Другим ријечима, имајући у виду овај примјер, улога претходно конфигурисаног, истестираног, и синхронизованог софтвера није тек тако безначајна, поготово када је ријеч о прављењу првих биоинформатичких корака.

БИО-ЛИНУКС КАО ПОМОЋНО НАСТАВНО СРЕДСТВО

Треба напоменути да је, поред тога што је требао да представља платформу за биоинформатичку анализу, сам Био-Линукс од почетка био замишљен и као својеврсно образовно помагало, за разлику од друге двије поменуте дистрибуције. Грубо речено, обезбијеђене су двије ствари:

1. Линукс-засновано радно окружење
2. конфигурисана биоинформатичка оруђа уз пратеће тестне податке.

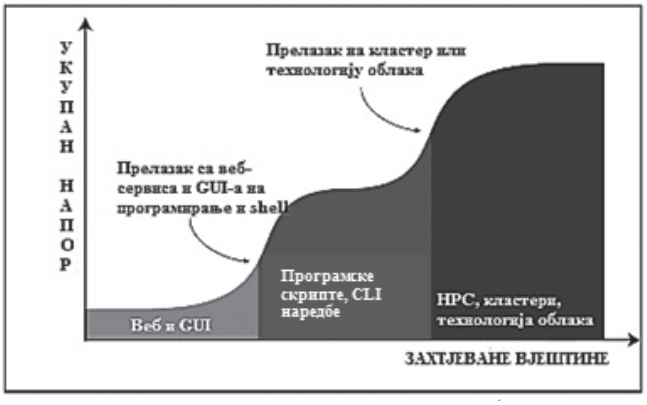
Ове двије ствари омогућују како савременом биологу-истраживачу, тако и студенту, да се усредсреди на

конкретне биоинформатичке проблеме, који су несумњиво исувише захтјевни да би се вријеме губило на нешто споредно. Иначе, у тзв. пост-геномичком истраживачком свијету UNIX-олике вјештине се веома цијене, што се лако може уочити прегледом уобичајених захтјева за запослење биоинформатичара на јавним огласима и порталима.

Према (7) постоје двије фазе – два стрма успона на кривој учења – у разоју бионформатичких вјештина за анализу СНП-података:

1. прелазак са интерактивних сесија на програмирање и развој биоинформатичких апликација
2. рад на умреженим/кластерованим рачунарским системима.

Оба прелаза захтијевају промјене у начину размишљања о самој анализи, укључујући и прилагођавање раду у новом рачунском окружењу.



Слика 4 *Два стрма успона на кривој учења током усвајања општих биоинформатичких вјештина.*

Кад је ријеч о првој фази, заиста је неопходно да се студентима омогући да уче у окружењу које је што је могуће сличније оном које ће користити у будућем истраживачком раду. Другим ријечима, једна образовна биоинформатичка платформа треба да садржи поред мноштва биоинформатичког софтвера и најпознатија програмска оруђа, или бар опцију за њихову аутоматску инсталацију:

- компилаторе за најпознатије програмске језике које користи биоинформатичка заједница: Перл, C/C++, Пајтон, R, BASH
- софтверске пакете за научна израчунавања
- развојна оруђа – редакторе (нпр. Nano, Gvim, Gedit, Kate), емулаторе текстовних љуски (нпр. Terminal, IPython), Make, интегрисана развојна окружења (нпр. Eclipse)

Улога готових тестних података, те примјера из праксе је очито немјерљива. Учитавање са флеш-драјва или пак DVD-а олакшава рад и учење у условима када није дозвољена инсталација новог оперативног система у рачунарским лабораторијама. Све у свему, можемо рећи да Био-Линукс јако добро прати трендове биоинформатичке заједнице и у стручном и у дидактичком смислу.

С друге стране, тј. кад је ријеч о другој фази, проблем је у томе што је број начина за похрану експерименталних података добијених данашњим методама секвенционирања просто „експлодирао". Са наглим развојем биотехнологије, самим тим и биоинформатике, јавља се и низ других проблема у вези са похрањивањем, сортирањем, претраживањем, али и коришћењем биолошких података. Подаци су буквално расути по својеврсним складиштима чији се број сваким даном увећава, при чему се поврх свега разликују и формати њиховог похрањивања. Штавише, неки од тих формата нису адекватни за аутоматску рачунарску обраду и морају проћи кроз тзв. препроцесирање, прије него што се искористе одговарајући алгоритми. Другим ријечима, развој биоинформатике је непосредно везан за такозвани „big data problem" – огромна количина хетерогених, сложених и недовољно структуираних података превазилази могућности традиционалне обраде (20). Интеграција података може донекле да ријеши неке од проблема који су у вези са big data феноменом. Комбиновање података из различитих извора и њихова заједничка анализа може да омогући комплекснији увид у разумијевање биолошких феномена, али је сам проблем пројектовања система за интеграцију података још увијек веома тежак. У циљу превазилажења поменутих потешкоћа аналитичари су суочени са два проблема: поставка и/или добијање приступа кластерованом рачунарском систему, и прилагођавање сопствених програмских скрипти и података за дјелотворан рад у новом окружењу. Но, неријетко процес добијања приступа за рад са кластерованим рачунарским системом не тече глатко, а јављају се и проблеми који се односе на рад са текстовном љуском, посебно прилагођеном за конкретан кластер. Штавише, чести су и оперативни пропусти који се односе на обезбјеђивање потребних програмских оруђа и биоинформатичког софтвера. На сву срећу, да тако кажемо, постоји алтернатива: технологија облака (енг. cloud computing).

С појавом поменуте технологије, рачун(ар)ски ресурси се могу закупити и прилагодити конкретним задацима, при чему се по завршетку њиховог извршавања дати ресурси остављају неком другом на употребу. У том погледу, рачунарски облак представља и веома корисно образовно помагало. Ипак, све што се закупи, тј. обезбиједи, као и сваки нов рачунар, долази у облику генеричких виртуелних машина, што захтјева прилагођавање рачунарског облака у сврсисходан систем. Због тога је својевремено покренут сестрински пројекат CloudBioLinux, настао из пројекта Био-Линукс, као свима доступна (виртуелна) групација машина за тзв. изразито захтјевне прорачуне у биоинформатици (енг. high-performance bioinformatics computing) (21). Технички, ClaudBioLinux је скуп одговарајућих програмских скрипти које омогућују поставку већ познатог окружења Био-Линукс, доступног за рад у облаку. Као што се сам оперативни систем може лако учитавати са флеш-драјва да би се обезбиједило конзистентно дидак-

тичко окружење, тако се и (лабораторијски) кластер може формирати на одговарајућем провајдеру (нпр. Amazon EC2) у само пар кликова мишем. Неизмијењена, стандардна конфигурација CloudBioLinux-а је већ спремна да задовољи широку палету савремених образовних захтјева. Због тога сматрамо да можемо с правом константовати да је Био-Линукс пријемчивија опција у односу на Scientific Linux и Fedora Scientiфic дистрибуције, које такође омогућују рад у облаку, али захтијевају не баш тако тривијална допунска подешавања.

УМЈЕСТО ЗАКЉУЧКА

Биоинформатика се као дисциплина веома брзо развија, па самим тим захтијева стално усавршавање и допуњавање знања. Савремене тенденције у биолошким истраживањима захтијевају све већу и већу употребу рачунарских система. Постоји стварна потреба да се превазиђе јаз између програмера с једне стране, и биолога с друге, као и да се међусобна сарадња учини што удобнијом. Штавише, постоје очигледне економске предности тог приступа. Што се тиче захтјева за идеалном софтверском платформом, испоставља се да су најближи у њиховом испуњењу бесплатни софтверски системи са јавно доступним кодом, поготово што је већина квалитетног софвера за биоинформатику развијена за тзв. POSIX-системе. Једно од предложених рјешења, које све више добија на популарности је оперативни систем Био-Линукс. Иако пројекат нема неку веома дугу историју, његов потенцијал је очит. Овакви и слични пројекти, чија су циљна група биолози-истраживачи, имају за циљ да продукују нову генерацију информатички образованих биолога, тј. биоинформатичара.

У раду су упоредно приказане карактеристике различитих био-оријентисаних дистрибуција на основу чега се лако долази до закључка да је Био-Линукс тренутно једина међу њима свеобухватна чисто био-оријентисана, а која у потпуности може да задовољи и стручне и дидактичке захтјеве биоинформатичке заједнице. Поред тога, поређење је проширено и на двије тренутно најпопуларније опште научно-оријентисане дистрибуције: Scientific Linux и FedoraScientific. Иако ове двије дистрибуције „боље котирају" у односу на Био-Линукс кад је ријеч о посљедњем ажурирању система, ипак лако се увиђа да је Био-Линукс тај који је de facto поставио одређене стандарде за све будуће био-оријентисане оперативне системе, поготово кад је ријеч о дидактичким аспектима и коришћењу технологије облака. Као што је већ поменуто, ове двије ствари захтијевају униформност чак и кад је ријеч о крајње техничким стварима, што је још један од разлога за оријентацију ка Био-Линуксу. Такође, дат је и примјер коришћења BLAST-а који се, додуше, могао извршити на све три поменуте дистрибуције, али је на посљедње двије била потребна допунска инсталација и прилагођавање софтвера. У суштини, увиђа се предност добијања готовог, ажурираног, и истестираног производа, са свим тестним подацима, који омогућује кориснику да се одмах усредсреди на конкретне биоинформатичке проблеме, који су као што смо већ казали, исувише захтјевни да би се вријеме губило на било шта споредно.

Што се тиче изазова из друштвене сфере биоинформатике, примјетан је мањак утрениpаних биоинформатичара не само код нас већ и у свијету. Могуће је да сама биоинформатика дјелује прилично тешко и конфузно. Једно од рјешења би могло да буде увођење интегрисаних студија. Тиме би и информатичари и биолози били много више упознати са постојећим софтверским рјешењима за биолошке задатке, радним биолошким платформама, а међу њима онда, наравно, своје мјесто би нашао и Био-Линукс. Такође, треба напоменути да се у региону нешто и дешава по овом питању. Сам Линукс није више толика енигма, штавише 4. јуна 2011. године у Бањој Луци се одржала и свјетска конференција Линукс/Дебијан програмера (22). С друге стране, у Београду (Република Србија), основан је Центар за биоинформатику, кога чине група научника, професора са Математичког факултета, а чији је рад у много чему значајан, те је и допринио развоју ове науке. Већина истраживања је била усмјерена на промјене структуре протеина садржаних у прокариотским ћелијама, а као крајњи резултат формирана је јавно доступна база података названа Prokaryote Disorder Database (23).

Имајући у виду све претходно речено, циљ овог рада је да упозна ширу, стручну заједницу са потенцијалом и могућностима Био-Линукса, те укаже на неке аспекте који би можда били од користи за развој биоинформатике код нас.

ЛИТЕРАТУРА

[1] Рачунарство у науци и образовању на почетку 21. века. Павловић-Лажетић, Гордана. 1-2, Београд : Математички институт САНУ, 2004, Настава математике, Т. XLIX, стр. 1-13.

[2] Bioinformatica: een werkconcept. Hesper, B and Hogeweg, P. 6, 1970, Kameleon, Vol. 1, pp. 28-29.

[3] The Roots of Bioinformatics in Theoretical Biology. Hogeweg, Paulien and Searls, David B. [ed.] David B. Searls. 3, 2011, PLoS Comput Biol., Vol. 7, p. e1002021. doi: 10.1371/journal.pcbi.1002021.

[4] Bioinformatics. Bayat, Ardeshir. 7344, 2002, BMJ: British Medical Journal, Vol. 324, pp. 1018-1022.

[5] What is the relevance of bioinformatics to pharmacology? Whittaker, PA. 2003, Trends Pharmacol Sci, Vol. 24, pp. 434-439.

[6] Parsing regulatory DNA: general tasks, techniques, and the PhyloGibbs approach. Siddharthan, R. 5, Aug 2007, J Biosci., Vol. 32, pp. 863-870.

[7] Bio-Linux as a Tool for bioinformatics Training. Booth, Timothy, et al., et al. Larnaca, Cyprus : s.n., 2012. Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE). pp. 578-582. DOI: 10.1109/BIBE.2012.6399736.

[8] BioFOSS: a survey of Free/Open Source Software in Bioinformatics. Shabaga, K and German, D. s.l. : IEEE, 2006. Computer-Based Medical Systems. DOI: 10.1109/CBMS.2006.60.

[9] How Perl Saved the Human Genome Project. Stein, L. 2, s.l. : The Perl Journal, February 1996, Vol. 1.

[10] Sequencing the SARS Virus. Krzywinski, M. November 2003, LINUX Journal, Vol. 115.

[11] Open software for biologists: from famine to feast. Field, Dawn, et al., et al. 7, July 2006, Nature Biotechnology, Vol. 24, pp. 801-803.
[12] Levine, Barry. Linux' 22th Birthday is Commemorated - Subtly - by Creator. CMSWire.com. [Online] Simpler Media Group, Inc., August 26, 2013. [Cited: August 19, 2016.] www.cmswire.com/cms/information-management/linux-22th-birthday-is-commemorated-subtly-by-creator-022244.php.
[13] Life sciences driven customized Linux distributions. Wajid, B and Serpedin, E. 1, January 18, 2014, OA Bioinformatics, Vol. 2.
[14] Linux distributions for bioinformatics: an update. Rana, A and Foscarini, F. 3, 2009, EMBnet.journal, Vol. 15.
[15] The bioinformatics playground. Tiwari, B and Field, D. 2005, Linux User & Developer, Vol. 46, pp. 50-56.
[16] Gelbmann, Matthias. Web Technologies of the Year 2015. W3Techs. [Online] January 4, 2016. [Cited: March 5, 2016.] http://w3techs.com/blog/entry/web_technologies_of_the_year_2015.
[17] Unsigned Integer Limited. Bio-Linux. DistroWatch.com. [Online] June 8, 2015. [Cited: May 8, 2016.] http://distrowatch.com/table.php?distribution=biolinux.
[18] Impacting the bioscience progress by backporting software for Bio-Linux. Paporovic, Sasa. arXiv, 2013, Computer Science. http://arxiv.org/abs/1310.1588v1.
[19] Basic local alignment search tool. Altschul, S, Gish, W, Miller, W i Myers, E, Lipman, D. 3, 1990, Journal of Molecular Biology, Tom. 215. doi: 10.1016/S0022-2836(05)80360-2.
[20] Adapting bioinformatics curricula for big data. Greene, Anna C., et al., et al. 2015, Brief Bioinform, pp. 1-8. doi: 10.1093/bib/bbv018.
[21] Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. Krampis, Konstantinos, et al., et al. 42, 2012, BMC Bioinformatics, Vol. 13, pp. 1-8. DOI: 10.1186/1471-2105-13-42.
[22] Министарство науке и технологије Републике Српске. Извјештај о ревизији финансијских извјештаја Министарства науке и технологије Републике Српск еза период 01.01.-31.12.20111. Бања Лука : Главна служба за ревизију јавног сектора Републике Српске, 2012.
[23] Bioinformatics analysis of disordered proteins in prokaryotes. Pavlović-Lažetić, Gordana M., et al., et al. 66, 2011, BMC Bioinformatics, Vol. 12, pp. 1-22. DOI: 10.1186/1471-2105-12-66.

Сретенка Видић, демонстратор на предмету Примјена рачунара у биологији, Универзитет у Бањој Луци, Природно-математички факултет
Контакт: sretenkavidic123k@gmail.com
Област интересовања: биоинформатика, биологија биљака

Дејан Кременовић, демонстратор на предмету Примјена рачунара у биологији, Универзитет у Бањој Луци, Природно-математички факултет
Контакт: dejan.kremenovic475@gmail.com
Област интересовања: биоинформатика, биологија биљака

Димитрије Д. Чвокић, асистент, Универзитет у Бањој Луци, Природно-математички факултет
Контакт: dimitriye.chwokitch@yahoo.com
Област интересовања: операциона истраживања, комбинаторна оптимизација, теорија игара, матхеуристике