DOI: 10.35191/MEDSOFT_2019_1_31_35_39

VÝZNAM A NÁROKY STRUKTURNÍ BIOLOGIE PRO VÝZKUM FUNGOVÁNÍ PROCESŮ V LIDSKÉM TĚLE A OBJEVOVÁNÍ LÉČIV

Tomáš Kulhánek

Abstrakt

Strukturní biologie se zabývá molekulární strukturou hlavně nukleových kyselin a proteinů a vlivu různých konformací na fyzikálně chemické vlastnosti. Rychlost a množství popsaných struktur je daleko za odhadovaným množstvím proteinů vyskytujících se v živých organismech a nové metody kombinují různé biofyzikální postupy.

Příspěvek shrnuje nároky a výzvy různých metod strukturní biologie na IT infrastrukturu a možnosti souvisejícího výzkumu jako je fyziologie člověka a objevování léku. Na příkladu vlastní zkušenosti z projektu pro tzv. integrativní strukturní biologii, jsou vyčteny metody pro sdílení výpočetních metod, dat, autentizaci, autorizaci a podpory řízení procesu experimentu pro jeho účastníky.

1 Úvod

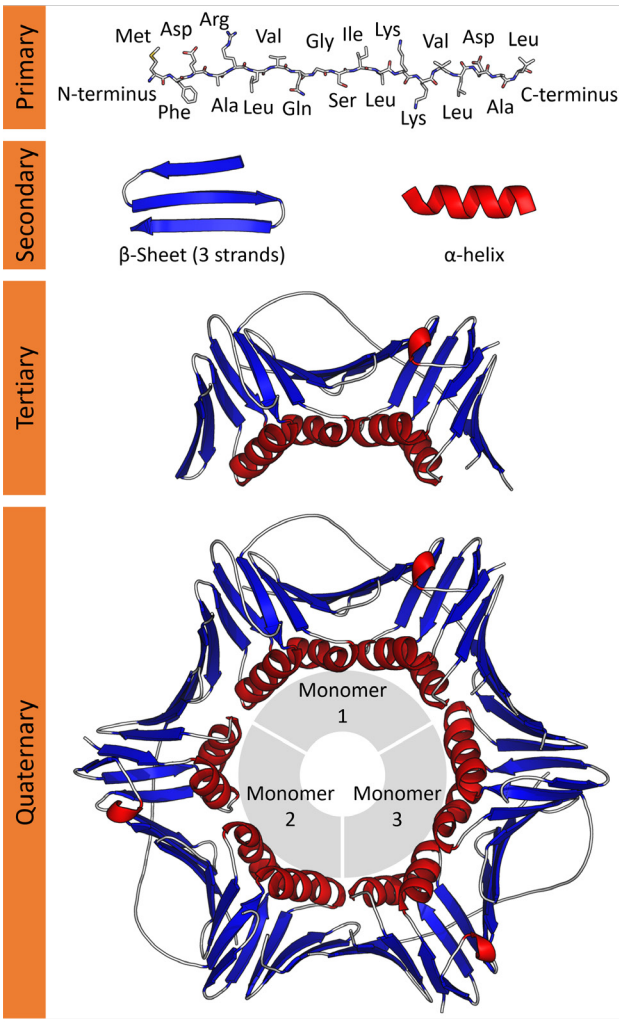
Strukturní biologie se zabývá molekulární strukturou látek v živých organismech, hlavně proteinů, nukleových kyselin a vlivu různých konformací na fyzikálně-chemické vlastnosti.

Primární struktura proteinu, tj. sekvence aminokyselin v řetězci bílkoviny je určena pořadím nukleových bází v genetickém kódu. V minulosti se věřilo, že rozluštěním genetického kódu se objasní struktura většiny proteinů, její vlastnosti a přeneseně vyřeší některé, či většinu otázek fyziologie a medicíny. Nicméně fyzikálně-chemické vlastnosti proteinu jsou určeny hlavně sekundární, terciární či quaternární strukturou (zjednodušeně jak se řetezec aminokyselin v bílkovině shluknе k sobě, aby vytvořil 3D strukturu a spojil se případně se stejnými nebo jinými bílkovinami do větších komplexů), viz obr. 1. Tuto 3D strukturu nelze z primární sekvence aminokyselin jednoduše odvodit. Struktura proteinů a nukleových kyselin se zjišťuje mnoha metodami od přímého zobrazování, jako jsou různé metody mikroskopie, tak nepřímých využívajících některé vlastnosti krystalů či magnetické rezonance a dalších. V současnosti je známa přesná struktura zhruba 130 000 proteinů, což je malý zlomek souboru proteinů vyskytující se v říši živých organismů. Nicméně tento počet narůstá každým rokem zhruba o 10% a vědecká komunita ukládá zjištěné strukturní informace do proteinových databází PDB, která je spravována třemi zakládajícími organizacemi RCSB v Americe, PDBe pro Evropu a Afriku (Protein Databank in Europe) a PDBj pro Asii (Protein Databank in Japan).[1], [2]

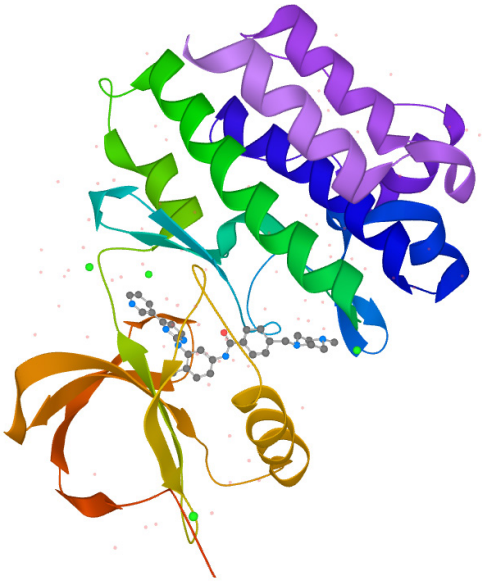
Strukturní informace se používá k návrhu léčiv, např. struktura malé molekuly STI-751 (Glevec®/Imatinib) byla navržena a ukázala se jako velmi specifická pro Abl kinázu vyskytující se v buňkách myelogenní leukémie a používá se jako léčivo pro tento a podobné typy nemocí. Viz obr. 2 a 3.[3] Strukturní znalost antigenu různých módů viru Meningokoka B umožnila navrhnout antigen, který se ukazuje jako kompletní ochrana proti různým typům viru meningokoka B. Tento objev otevřel novou doménu racionálního návrhu vakcín, tzv. strukturní vakcinologie. [4]–[6]

2 Metody strukturní biologie

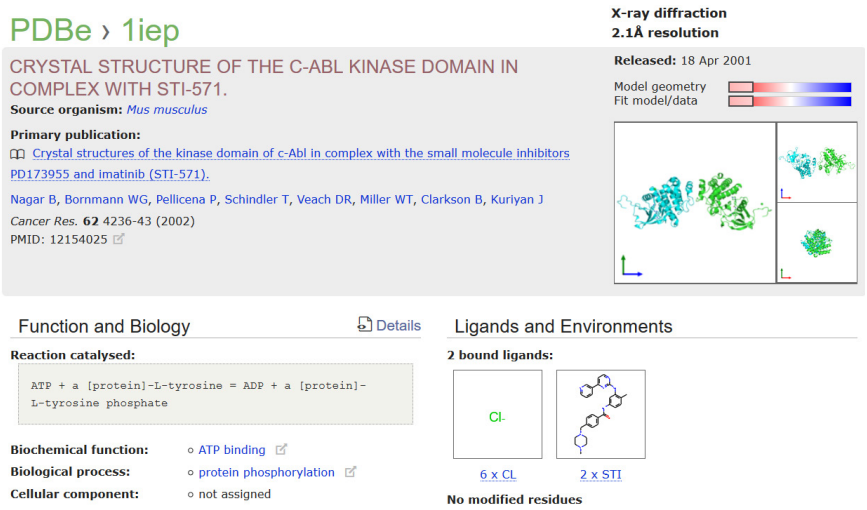
Hlavními metodami strukturní biologie jsou především rentgenová krystalografie (X-Ray crystallography, 90% záznamů v PDB databázi), nukleární magnetická rezonance (NMR, 10% záznamů v PDB databázi) a kryoelektronová mikroskopie (Cryo-EM, 1% záznamů v PDB databázi).



Obrázek 1 – Struktura proteinu od primární do quaternární složená z 3 monomerů. Příklad proteinu proliferating cell nuclear antigen (PCNA) v PDB databázi uložené pod záznamem 1AXC. Zdroj: Thomas Shafee, https://en.wikipedia.org

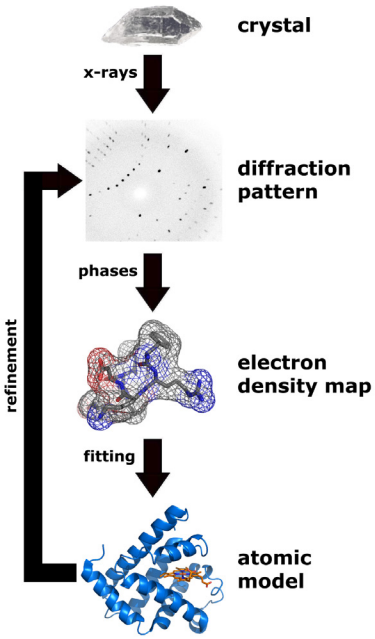


Obrázek 2 – Struktura komplexu C-ABL kinázy s STI-571, PDB záznam 1IEP

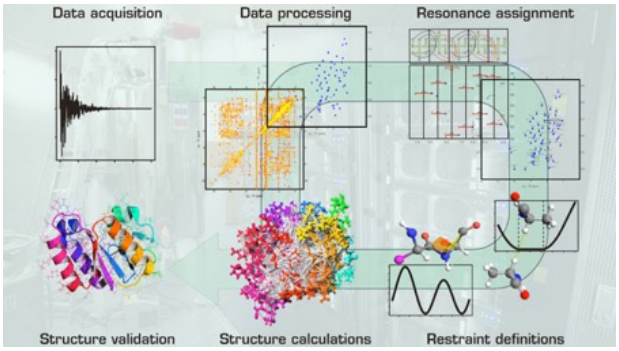


Obrázek 3 – PDB záznam 1IEP pro komplex C-ABL kinázy, odkazy na původní článek, vizualizace a další detailní informace poskytované databází PDBe v Evropě spravované EMBL-EBI v Cambridge.

Rentgenová krystalografie (X-Ray crystallography) využívá jevu difrakce, který v krystalické struktuře vytváří na snímaném obrazu tzv. vzory (diffraction patterns). Pomocí několika stovek až tisíc snímků z různých úhlů lze výpočetními metodami rekonstruovat tzv. elektronovou hustotu a rekonstruovat strukturu podle odhadované primární struktury proteinu – modelu (obr. 4). Zdrojem rentgenového záření jsou v dnešní době hlavně synchrotrony a lasery. Tato zařízení jsou poměrně finančně náročná na stavbu a provoz, proto je budují a provozují buď velké státy (Francie, Anglie,Německo, ...), nebo mezinárodní konsorcia (CERN). Během experimentu se pořídí obrazová data o velikosti několika GB a balíky analytického software většinu kroků při zpracování dat automatizují. Zpracování dat se odehrává většinou na místě, neboť tato velká vědecká centra jsou vybavena nadstandardní výpočetní a úložnou kapacitou, jednou z největších ve vědecké komunitě.



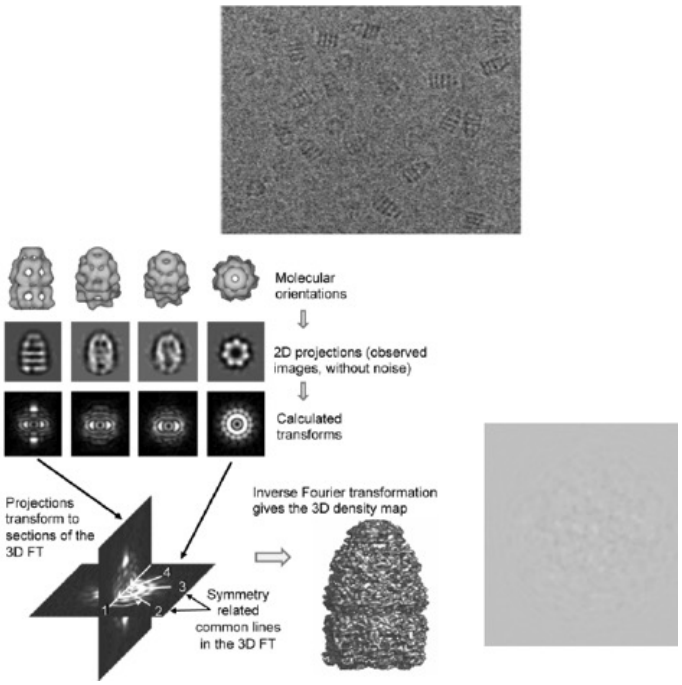
Obrázek 4 – Tok vědeckých dat v rentgenové krystalografii. Svazek rentgenových paprsků prosvítí krystal, na snímku se projeví jako difrakční vzor. Z mnoha snímků lze postupně rekonstruovat mapu elektronové hustoty a napasovat model na úrovni atomů.



Obrázek 5 – Hrubě zjednodušené znázornění zpracování dat z nukleární magnetické rezonance. Silný magnetický puls rezonuje se vzorkem a jsou detekovány frekvence typické pro některá jádra atomů a molekul. V dalším zpracování dat je možné odvodit relativní vzdálenosti jader a celý proces se opakuje pro různé frekvence pulzů a dalších veličin. Lze tak postupně rekonstruovat strukturu zkoumané látky.

Nukleární magnetická rezonance detekuje, jak daný vzorek absorbuje radiofrekvenční signály v silném magnetickém poli a postupně lze tyto informace použít k stanovení vzdáleností mezi jádry různých atomů a molekul ve vzorku a přeneseně i struktury. Zařízení pro NMR jsou menší a mohou si je dovolit univerzity nebo meziuniverzitní centra. Balíky analytického softwaru pak automatizují některé kroky při zpracování naměřených spekter spolu s daty z veřejných databází. Data vyprodukována NMR metodami se pohybují řádově v desítkách až stovkách MB. Nicméně zpracování je velmi náročné na výpočetní kapacitu a proto velké množství výpočtů a algoritmů je poskytováno jako služby v gridové a cloudové infrastruktuře např. WeNMR ve vědeckých komunitách (např. INFN v Itálii, nebo mezinárodní konsorcium EGI). (Obr. 5)

V poslední době probíhá tzv. revoluce v rozlišení (resolution revolution) v oblasti kryoelektronové mikroskopie. Tato metoda díky sofistikovanému zpracování velkého množství pořízených snímků se dostává z rozmazané předlohy k rozlišovacím schopnostem rentgenové krystalografie. Vzorky jsou hluboce zmrazené a poté jsou nasnímány elektronovým mikroskopem z mnoha různých pozic a úhlů. Výsledná sekvence, dá se říct videosekvence, je zpracována výpočetními algoritmy a rozmazání je eliminováno a dokáže eliminovat



Obrázek 6 – Cryo-EM, kryo elektronová mikroskopie, z mnoha projekcí téže molekulární struktury se sestaví 3D mapa

šum a zvýšit rozlišení. Podobné metody se používají např. v astronomii, kdy se dotyčný astronomický objek snímá po delší dobu včetně mihotání vzduchu, které obraz náhodně rozostřuje, z dostatečně dlouhé videosekvence lze rekonstruovat poměrně dobrý obraz s detaily, které jsou nad rozlišovací schopností daného přístroje a v daných podmínkách. (Obr. 6)

3 Výpočetní a datové nároky pro hybridní, integrativní přístup

Každá z metod má své limity a výhody. Nicméně poslední dobou se ukazuje, že kombinací dvou a více metod při studiu proteinů je výhodné neboť např. ve fázi refinement všech výše zmíněných přístupů se využívají validované struktury podobných proteinů či částí proteinů pořízených pomocí jiných metod, přičemž ne vždy byl příslušný experiment proveden na daném vzorku.

Každý z výše zmíněných přístupů používá vlastní pravidla pro nakládání s daty. Např. velké synchrotrony uchovávají data po dobu 30 dnů, kdy si je autor nebo autoři experimentu mohou nahrát buď na vlastní disky nebo přenést dle vlastního uvážení do dlouhodobějšího archivu. Tento přístup ponechaný na uvážení vědce vede po nějaké době často ke ztrátě nebo nedostupnosti původních dat[7] a proto vznikají iniciativy pro uchovávání hrubých dat dlouhodobě, datové archivy pro hrubá data např. Zenodo apod. Menší centra a laboratoře obvykle mají ad-hoc pravidla pro nakládání s daty a pro hybridní metody je důležité, aby se data dala přenést, použít, zpracovat na původně nezamýšleném místě. S tímto ohledem se vyvíjí např. služby pro agregaci dat, nebo integraci různých datových poskytovatelů.

Každá metoda má taktéž vlastní sadu analytického softwaru ať už licencovaného nebo otevřeného. GROMACS a AMBER jsou balíky programů pro simulaci molekulární dynamiky, které vyhodnocují síly působící na všechny atomy ve studované molekule a aktualizují rychlost pohybu a pozici atomů podle Newtonových pohybových rovnic a generují termodynamické chování systému, či spočítají hodnoty volné energie molekul. Tyto programy se používají při fázi vylepšování (refinement) podle experimentálních dat výše zmíněných metod [8], [9]. HADDOCK je balík programů a služeb pro simulování a modelování spojení více makromolekul, které mohou být použity pro predikci a ladění dalších experimentů.[10] RELION je program využívající některých vlastností snímků kryo elektronové mikroskopie a statistických metod pro strukturní rekonstrukci a při minimální intervenci uživatele [11].

Krom klasické distribuce softwaru formou balíků, které si konečný uživatel nainstaluje na svém počítači a provozuje, se dnes poskytuje software formou služby a přístupu ke sdíleným výpočetním zdrojům v gridu nebo v cloudu. V Česku pro akademickou a medicínskou komunitu infrastrukturu poskytuje např. sdružení CESNET[12], [13]. V mezinárodním prostředí buď národní poskytovatelé sloučení do organizace EGI, jehož CESNET je členem, nebo výpočetní centra vědeckých center (CERN, STFC, …), které část své kapacity sdílejí formou gridu či cloudu pro vědeckou komunitu. Údržba doménově specifických balíků je pak na vědecké komunitě, např. zmíněné balíky HADDOCK a GROMACS jsou mj. obohacovány o další aplikační rozhraní pro webové či jiné simulace v projektu Bioexcel. AMBER je v gridové komunitě přístupný přes webové rozhraní a dostupný pro NMR komunitu[14]. RELION je součástí větších balíků např. SCIPION s možností individuálního výpočtu na dedikované virtuálním stroji podle SCIPION Cloud

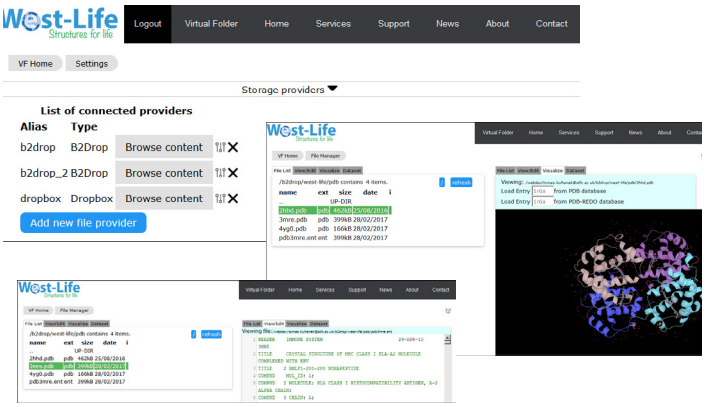
Při použití hybridních metod, je zpracování dat pomocí softwaru náročnější, neboť vyžaduje propojení softwarových balíků a dat z různých uložišť, která nebyla původně zamýšlena pro takové spojení. Uživatel, většinou Ph.D. student nebo postdoc je odborník v jedné experimentální doméně není obvykle zběhlý v softwaru pro jinou doménu (nemluvě o instalaci a údržbu softwarových balíků). Z výše vyjmenovaných nároků při zpracování dat ve strukturální biologii plynou tyto požadavky na současné a budoucí systémy:

1. zpracovávat data z různých zdrojů. Integrovat datová úložiště pomocí dostupných a podporovaných technologií.
2. používat různorodý software a webové služby
3. umožnit nově vznikajícím službám jednoduše používat sdílená data, software a webové služby
4. autentikace a autorizace pro řízení přístupu ke zdrojům, tj. např.
   umožnit přihlášení pomocí existujících institucionálních či veřejných účtů

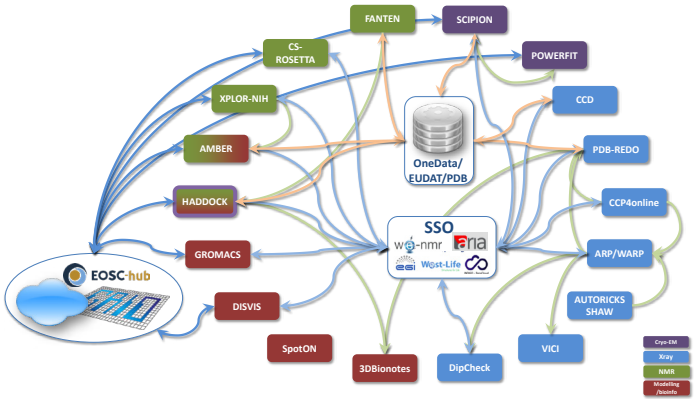
V CERNu si uvědomili problematiku distribuce různých softwarů a konfigurací, proto již několik let vyvíjejí obecnou technologii CernVM-FS a virtuální operační systém CernVM. Konečný uživatel si může spustit univerzální virtuální stroj a poté zvolit kontext, ve kterém chce pracovat (tj. jaký konkrétní softwarový balík a konfigurace jsou mu v rámci virtuálního stroje dostupné) [15], [16] Oproti klasické kontextualizaci ve službách Amazon AWS, MS Azure, Google Cloud jsou využívané síťové prvky a cache pro výkonnou distribuci softwaru a dat přímo ke koncovému virtuálnímu stroji. Díky tomu lze dosáhnout zajímavých časů pro první naběhnutí systému, nebo konkrétní aplikace.

Na výše zmíněné požadavky cílí projekt West-Life v doméně strukturální biologie. Jeho Virtuální složka (virtual folder, obr. 7) umožňuje registrovat uživateli jeho datová úložiště roztroušená po různých místech a službách (např. komerční DROPBOX, nebo vědecký B2DROP přes rozhraní WEBDAV) a poté např. vygenerovat přístupový bod (URL) pro existující nebo novou webovou službu, která data zpracuje nebo uloží. Přístup k datům je transparentní na úrovni souborového

systému, takže např. virtuální stroj může k registrovaným datovým úložištím přistupovat přes souborový systém [17]. Díky integraci tzv. single-sign on se uživatel ke službě přihlašuje institucionálním účtem, nebo účtem existujících služeb jako je ARIA Instruct nebo WeNMR. Pro náročnější uživatele je pak připraven obraz virtuálního stroje, který si může spustit na vlastním či pronajatém hardware a který v sobě zahrnuje integrace datových úložišť ve formě virtuální složky a přístup k běžným softwarovým balíkům pro strukturální biologii (obr. 8) distribuovaným pomocí technologie CernVM-FS a repozitářům udržovaným v rámci EGI organizací STFC [18].



Obrázek 7 – Uživatelské rozhraní virtuální složky (virtual folder) projektu West-Life. Uživatel registroval své účty s daty u poskytovatelů Dropbox a EUDAT B2DROP. Jednoduchý souborový manažer může prohlížet adresáře a soubory. Soubory formátu PDB nebo mmCIF jsou vizualizovány pomocí widgetu Litemol.



Obrázek 8 – Služby a integrace v projektu West-Life. Zelené propojení (např. HADDOCK, SSO) označuje přímou integraci software/služeb. Růžové (např. HADDOCK, OneData/EUDAT/PDB) označuje propojení přes datové uložiště. Modré označuje interakci s některou s přihlašovacích služeb (SSO). Navíc jsou znázorněny služby přístupné přes EOSC-Hub, středisko vědeckých služeb pro iniciativu otevřeného vědeckého cloudu – EOSC (european open science cloud).

4 Důsledky pro ostatní obory

Výše zmíněné příklady výsledků softwarových systémů a webových služeb v oboru strukturní biologie lze aplikovat i v jiných oborech, neboť použité technologie a postupy jsou obecné.

Sdílení dat a sdílení přístupu k datovým úložištím umožňuje integraci softwaru a služeb na té jednodušší úrovni. Příklad virtuální složky (virtual folder) ukazuje jak pro program, který umí číst a zapisovat pouze do souboru, lze snadno integrovat do služeb typu DROPBOX, B2DROP apod. Ne všechna velká vědecká centra povolují přístup k datům pomocí aplikačního rozhraní třetích stran nebo protokolů (WEBDAV, apod.), v takovém případě je na uživateli, aby si přenesl svá data k dlouhodobějšímu uložení např. K službám EUDAT B2DROP aj.

Obraz virtuálního stroje s kontextem lze využít k rychlé instalaci a zpřístupnění celého výzkumného prostředí se specializovaným softwarem, technologie VmWare, VirtualBox, Docker. Integrace se SSO pak přináší podporu autentikace a autorizace bez nutnosti technicky zajišťovat a udržovat databázi uživatelů, ale zároveň být v souladu např. s GDPR

Mezi týmy zkoumajícími podobnou oblast, vyvíjejícími podobné nástroje a metody je obvykle jistá míra rivality. Příklad zmíněných projektů strukturní biologie je také příkladem, jak rivalita mezi týmy byla naopak využita pro spojení různých metod a služeb k obohacení komplexnějšího výzkumu přesahující jednooborové zaměření.

Problematikou dlouhodobého uchovávání hrubých dat, anotace a vyhledávatelnosti a reprodukovatelnosti výsledků se zabývá v současnosti řada projektů. EUDAT jako projekt spojil nejprve existující infrastrukturu několika partnerů. Dnes už čítá 25 partnerů po celé Evropě a nabízí služby spojené s uchováváním a vyhledáváním dat. Současné programy a softwarové balíky za několik let a desetiletí zastarávají. Krom ztráty dat se může ztratit schopnost data přečíst a zpracovat, např. kvůli zastarávání formátu, nebo kvůli ztracené, zastaralé verzi výpočetního software. Kromě hrubých dat vznikají iniciativy a standardy pro uchování původu dat, tj. kroků a vazeb mezi daty a dalšími agenty, které se mohou použít pro znovurealizaci výpočtu s konkrétními parametry (např. W3C standard PROV-O).

V současnosti obvyklá 3 vrstvá architektura webových aplikací (klient-server-databáze) je v některých případech nahrazována architekturou server-less (backend-less, unhosted) kde autor aplikace neprovozuje server, server a databázi používá a pronajímá si pomocí služeb PaaS, případně výběr kompatibilních serverů nebo úložišť dat nechává plně na uživateli. Toto paradigma umožňuje už v současnosti nabízet tzv. Widgety (např. vizualizační Litemol[19]), které dávají přidanou hodnotu k diametrálně odlišným aplikacím viz např. Virtuální složka (obr. 7).

Literatura

[1.] *P. W. Rose et al., "The RCSB protein data bank: Integrative view of protein, gene and 3D structural information," Nucleic Acids Res., vol. 45, no. D1, pp. D271–D281, Jan. 2017.*

[2.] *S. Velankar et al., "PDBe: Improved accessibility of macromolecular structure data from PDB and EMDB," Nucleic Acids Res., vol. 44, no. D1, pp. D385–D395, Jan. 2016.*

[3.] *M. Deininger, E. Buchdunger, B. J. Druker, and J. V Melo, "The development of imatinib as a therapeutic agent for chronic myeloid leukemia.," Blood, vol. 105, no. 7, pp. 2640–53, Apr. 2005.*

[4.] *H. Tettelin et al., "Complete genome sequence of Neisseria meningitidis serogroup B strain MC58.," Science, vol. 287, no. 5459, pp. 1809–15, Mar. 2000.*

[5.] *R. Rappuoli, "Reverse vaccinology, a genome-based approach to vaccine development," Vaccine, vol. 19, no. 17–19, pp. 2688–2691, Mar. 2001.*

[6.] *R. Cozzi, M. Scarselli, and I. Ferlenghi, "Structural vaccinology: a three-dimensional view for vaccine development.," Curr. Top. Med. Chem., vol. 13, no. 20, pp. 2629–37, 2013.*

[7.] *P. Bryan Heidorn, "Shedding Light on the Dark Data in the Long Tail of Science," Libr. Trends, vol. 57, no. 2, pp. 280–299, 2008.*

[8.] *D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and*

H. J. C. Berendsen, "GROMACS: Fast, flexible, and free," Journal of Computational Chemistry. 2005.

[9.] R. Salomon-Ferrer, D. A. Case, and R. C. Walker, "An overview of the Amber biomolecular simulation package," Wiley Interdiscip. Rev. Comput. Mol. Sci., 2013.

[10.] G. C. P. Van Zundert et al., "The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes," J. Mol. Biol., 2016.

[11.] S. H. W. Scheres, "RELION: Implementation of a Bayesian approach to cryo-EM structure determination," J. Struct. Biol., vol. 180, no. 3, pp. 519–530, 2012.

[12.] T. Kulhánek, M. Mateják, J. Šilar, and J. Kofránek, "IDENTIFIKACE FYZIOLOGICKÝCH SYSTÉMŮ," in sborník příspěvků MEDSOFT, 2014, pp. 148–153.

[13.] J. Navrátil, S. Ubik, R. Igliar, and P. Pečíva, "CESNET A JEHO AKTIVITY V MEDICÍNĚ," in Medsoft, 2015, pp. 157–169.

[14.] I. Bertini, D. A. Case, L. Ferella, A. Giachetti, and A. Rosato, "A Grid-enabled web portal for NMR structure refinement with AMBER," Bioinformatics, vol. 27, no. 17, pp. 2384–2390, Sep. 2011.

[15.] J. Blomer, G. Ganis, N. Hardi, and R. Popescu, "Delivering LHC Software to HPC Compute Elements with CernVM-FS," Springer, Cham, 2017, pp. 724–730.

[16.] J. Blomer et al., "Micro-CernVM: slashing the cost of building and deploying virtual machines," J. Phys. Conf. Ser, vol. 513.

[17.] T. Kulhanek, C. Morris, and M. D. Winn, "West-Life Virtual Folder—Components and Tools," in Instruct Biennial, 2017.

[18.] Morris, Chris et al., "West-Life: a virtual research environment for structural biology," J. Struct. Biol. X, vol. in press, 2019.

[19.] D. Sehnal et al., "LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data," Nat. Methods, 2017.

Kontakt

Tomáš Kulhánek
Science and Technology Facility Council
United Kingdom
e-Science Data Factory,
France