Bioinformatika

Bioinformatika je interdisciplinární obor, který vyvíjí a používá nástroje, které usnadňují pochopení komplexních medicínských dat. Jedná se o přírodní vědu jako každá jiná, akorát je v porovnání s například botanikou velice mladá. Jelikož bioinformatika v poslední době rapidně nabývá na významu, lékařské fakulty (zejména ty lepší) vyčleňují na její výuku větší časové bloky a samostatné ústavy.

Bioinformatika se stává nezbytnou součástí vědecké práce, jelikož všechny nové poznatky a data se začínají skladovat a zpřístupňovat digitálně a pro širší využitelnost a větší trvanlivost starších analogových materiálů se i ony začínají převádět do digitální podoby. Některé informace, jako například databáze organických sloučenin nebo genetický kód navíc nabývají takových velikostí, že jejich skladování v analogové podobě se stává nemyslitelným. Bioinformatika dále umožňuje lepší vzájemné propojení jednotlivých vědních oborů a usnadňuje spolupráci vědců z různých oblastí výzkumu.

Největší význam dnes bioinformatika nachází v:

- Genetice (nejnovější sekvence lidského genomu má například přes 3 miliardy nukleotidů)
- Analýze mikroskopických obrazů (například při počítání buněk určitého typu v preparátech)
- Molekulární biologii (významně například v proteomice kde se využívá k modelaci proteinů a zkoumání jejich struktury)
- Farmacii (predikce funkce a účinků jednotlivých molekul)

Data

Základním objektem bioinformatiky jsou **data**.

> *Data je výraz pro údaje používané pro popis nějakého jevu nebo vlastnosti pozorovaného objektu.* – Wikipedia

Daty mohou být například obrázky z mikroskopu, chemické vzorce, sekvence nukleotidů nebo statistika počtu pacientů s nějakou nemocí v čase. Bioinformatika se zabývá zpracováváním těchto dat a jejich prezentací.

Data mají omezenou přesnost – například rozlišení obrázku nebo perioda zaznamenávání počtu pacientů. Přístroje generující data se však neustále zlepšují a dat tak exponenciálně přibývá. V situaci, kdy z jednoho mikroskopu můžeme získat každý den statisíce obrázků, není myslitelné, že bychom je vyhodnocovali ručně. Ruční analýza dat má další nevýhodu, a to chyby způsobené lidským faktorem.

Dalším příkladem obrovského objemu dat jsou lidské genomy. Počet sekvenovaných genomů je každý rok větší z důvodu exponenciálního snižování ceny sekvenace v průběhu času. Nejlepší sekvenátory vyprodukují po celém světě 148 TB dat každou hodinu. To je více, než data zapsaná ve všech knihách, co kdy lidstvo napsalo, dohromady.

Pokud máme data uložena v digitálním formátu, můžeme na nich mnohem snáze provádět různé operace, jako vyhledávání, odhalování trendů nebo analýza.

Analýza dat

Analýza nasbíraných dat se skládá ze tří kroků:

1. **Příprava**
2. **Vlastní analýza**, která se skládá z:
   - Importu
   - Očištění
   - Pochopení
   - Komunikaci výsledků
3. **Sdílení**

Z těchto kroků paradoxně často zabírá nejvíce času příprava. Analýzu ponecháváme strojům, a tak tento automatizovaný krok často probíhá velice efektivně i při velmi velkých objemech dat.

Strojové učení

Strojové učení je proces, kterým učíme stroje zpracovávat nebo analyzovat data takovým způsobem, který potřebujeme. Rozlišujeme několik hlavních typů, které se hodí pro různé úkony.

1. **Bez učitele** (*unsupervised*)
   - **Shlukování** (*clustering*) – Agregace dat do skupin podle objevených podobností (například CD markery na leukocytech)
   - **Redukce dimenzionality** (*dimensionality reduction*) – Zobrazení dat vypovídajících o mnoha vlastností do menšího počtu dimenzí (často 2 - 3) nahrazením původních proměnných menším množstvím nových, které je kombinují (například graf závislosti poměru SYS a DIA tlaku na příčině smrti a věku dožití)

2. **S učitelem** (*supervised*)
   - **Regrese** (*regression*) – Ustanovení modelu z pozorovaných dat a predikce závislé proměnné z nezávislé podle něj (například ustanovení šance na úmrtí na srdeční selhání podle koncentrace lipoproteinů v krvi)
   - **Klasifikace** (*classification*) – Rozřazení objektů do pouhých několika skupin podle mnoha jejich vlastností (například rozhodnutí, zda má být příchozí e-mail označen jako spam)

Všechny z těchto metod se hodí na více věcí, stejně tak pro provedení některého typu analýzy můžeme využít více z těchto metod.

Zpracování dat

Bioinformatika v sobě zahrnuje mnoho ostatních oborů, nicméně vždy zahrnuje **bio**logii a **informatiku**. Biologie je zásadní pro pochopení toho, jaká data potřebujeme, jak máme interpretovat výsledky a jaké mezi nimi existují vztahy (například sekvence DNA a proteinová skladba buňky). Informatika je zase nezbytná pro zobrazení sesbíraných biologických dat v často obskurních formátech a k jejich automatizovanému zpracování a interpretaci.

Podle toho, jak specifické úkony chceme na svých datech provádět můžeme využívat čtyři různé úrovně softwaru.

1. Uživatelsky přívětivé programy a webové aplikace – jsou obvykle graficky příjemně nadesignované a "blbuvzdorné", nicméně jejich vývoj je zdlouhavější a existují tak obvykle pouze pro běžnější datové úkony.
2. Hotové programy vyžadující ovládání z příkazového řádku – jsou k dispozici i pro méně obvyklé úkony díky tomu, že není potřeba při vývoji ztrácet čas s tvorbou UI a lze je tak vyrábět rychleji a ve větších množstvích
3. Kombinování nástrojů do analytického protokolu
4. Vývoj vlastních nástrojů – Pokud potřebujeme udělat velice specifický úkon na našich datech, může se stát, že budeme první na světě, kdo jej bude potřebovat a nebudeme mít k dispozici žádné programy. V takové situaci si musíme nástroj pro analýzu vytvořit sami. V okamžiku, kdy se staneme touto cestou vývojáři, můžeme si přesně vybrat, jakým stylem budou stroje naše data zpracovávat a jak je budou zobrazovat. Stejně tak můžeme vytvořit i grafické rozhraní a umožnit tak používání našeho softwaru širšímu okruhu vědců.

Programování

Program není nic jiného než seznam elementárních úkonů, které má počítač provést. Jednotlivé příkazy se mohou opakovat, provádět se pouze za námi definovaných podmínek, nebo předčasně ukončit běh programu. V bioinformatice se v současné době nejčastěji používají programovací jazyky **Python** a **R**.

Programy se vyvíjejí ve **vývojových prostředích** (IDE). Jedná se o "vylepšené" textové editory, které nám usnadňují psaní kódu našeptáváním příkazů a odhalování syntaktických chyb. Další nezbytnou součástí programátorského kufříku je **verzovací software** (z nichž nejznámější je Git (https://git-scm.com)), který nám umožňuje vracet se k předchozím verzím našeho skriptu. Velice důležitá je i **dokumentace**, která nám zajišťuje, že se ve vlastním kódu neztratíme a v ideálním případě také to, že jej po nás pochopí i někdo jiný a bude jej moci například rozšířit. Pro spolupráci s ostatními programátory/bioinformatiky a sdílení kódu se využívají **online repozitáře**, například GitHub (https://github.com) nebo BitBucket (https://bitbucket.org).

Buzz Words související s bioinformatikou

- **Big data** – nejasná definice, používá se obvykle pro velké datasety, pro jejichž ukládání je potřeba zvláštní infrastruktura
- **Artificial intelligence (AI)** = umělá inteligence – program, který je schopen reagovat na širokou variaci uživatelských vstupů, například chatboti na bankovních stránkách; mohou být i velice složité a věrně imitovat myšlení skutečného člověka
- **Data mining** – získávání dat skenováním sociálních sítí, záznamů pacientů nebo třeba historie zpoždění vlaků a jejich ukládání v konzistentním formátu
- **Machine learning** = strojové učení – proces, kterým učíme počítače přicházet s řešením našich problémů namísto toho, abychom jim návod k řešení poskytli sami
- **Cloud computing** – přesun výpočetní práce z našeho lokálního zařízení na server (do cloudu), který obvykle disponuje větší výpočetní silou (například infrastruktura CESNET (https://www.cesnet.cz))
- **Virtualizace** – vytvoření testovacího prostředí imitujícího samostatný systém uvnitř výkonnějšího zařízení

FISHER, Karel. *Bioinformatika* [přednáška k předmětu Metodologie vědy a bioinformatika, obor Všeobecné lékařství, 2. lékařská fakulta Univerzita Karlova]. Praha. 30. 11. 2022 11:40–13:20. Profil přednášejícího ve WHOIS systému UK (https://is.cuni.cz/webapps/whois2/osoba/1947528388653157).