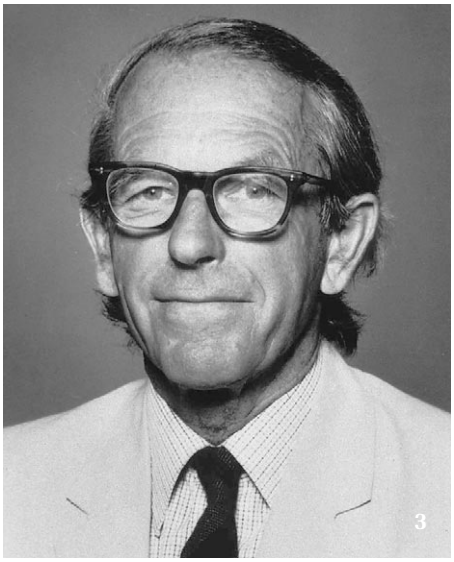
Marian Novotný

Revoluce, ne evoluce – jak se neutopit v záplavě biologických dat

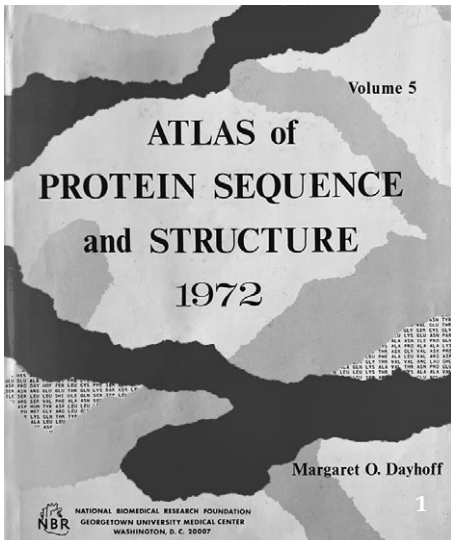


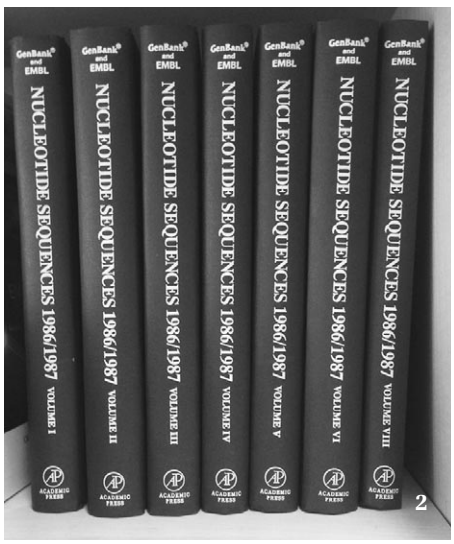
1 a 2 Ukázky databází v knižní formě, jak byly distribuovány v předinternetové době. První proteinová databáze od Margarety O. Dayhoffové vyšla v r. 1965 (na obr. 1 vydání z r. 1972) a několik vydání první velké nukleotidové databáze Genbank a EMBL (European Molecular Biology Laboratory, 2). Foto G. Barton (https://X.com/gjbarton; obr. 1) a D. Landsman (Wikimedia Commons, 2)
3 Frederick Sanger, autor sekvenačních metod pro DNA, RNA i proteiny
4 M. O. Dayhoffová, autorka první proteinové databáze. Převzato z Wikimedia Commons (obr. 3 a 4)

Biologové vždy rádi sbírali data. I mezi čtenáři Živy je jistě mnoho takových, kteří mají doma entomologickou síťku a sbírku hmyzu nebo strávili mnoho času sběrem a přípravou materiálu pro herbář. U královských či šlechtických dvorů byly často provozovány zvěřince, a to už ve starověkém Egyptě, u nás měl na Pražském hradě zvěřinec např. Rudolf II. Nejstarší stále provozovaná zoologická zahrada se nachází ve Vídni a byla založena v r. 1752. Nejstarší botanická zahrada – Orto Botanico di Padova – byla založena Benátskou republikou v Padově již r. 1545 k pěstování léčivých rostlin, ale také jako místo, které mělo učit rozpoznávat nebezpečné rostliny. S objevem mikroskopu vzrostl zájem o organismy a části jejich těl, které jsou příliš malé pro pozorování lidským okem, a následně i o sběr dat, popisujících tyto organismy. V druhé polovině 20. století se vědci postupně naučili získávat další typy dat, a to především k molekulárním detailům života.

Díky talentu a úsilí Fredericka Sangera (obr. 3) umíme popsat pořadí (sekvenci) základních stavebních kamenů DNA, RNA i proteinů. Zásluhou dalších osobností, jako byl John C. Kendew a Max F. Perutz, umíme určit 3D strukturu makromolekul. Dále dokážeme identifikovat vazebné partnery makromolekul, popsat kompletní genetickou informaci organismu, zjistit rozdíly v množství dostupného proteinu mezi zdravým a nemocným člověkem, automatizovat sběr mikroskopických snímků a aplikovat mnoho dalších experimentálních metod.

Sběr molekulárních dat byl na začátku extrémně pracný, pomalý a mimořádně drahý. Tato data bývala generována k zodpovězení specifické otázky, ale mohou být později použita i k řešení jiných, často mnohem obecnějších otázek, a proto jsou po skončení experimentu ukládána do databází.





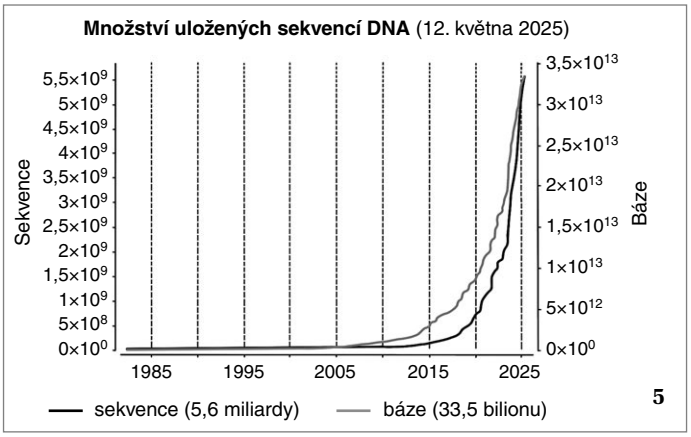
Databáze slouží jako snadno dostupný zdroj informací, který je optimalizován pro efektivní vyhledávání. Ukládání do databází brání ztrátě generovaných dat a umožňuje i kontrolu jejich kvality. Databáze molekulárních dat jsou mnohem mladší než např. botanické zahrady a jejich vznik i rozvoj byl závislý na pokroku v oblasti informačních technologií. První databázi proteinových sekvencí založila v r. 1965 americká biofyzička Margaret Dayhoffová – jmenovala se Atlas of Protein Sequence and Structure (obr. 1 a 4). Vyšla jako kniha a obsahovala všechny v té době známé sekvence proteinů – shodou okolností jich bylo 65. Kniha se následně r. 1984 transformovala v první on-line databázi PIR (Protein Information Resource), která byla dostupná po telefonních linkách a přístupná pro vzdálené počítače. První databáze 3D struktur makromolekul, zvaná PDB (Protein Data Bank), byla založena v r. 1971 a obsahovala 7 struktur. Databáze nukleových kyselin Genbank vznikla v r. 1982 a zahrnovala okolo 2 000 sekvencí.

Velikost databází při vzniku dokazuje skromné začátky sběru molekulárních dat, ale málokterá oblast lidské činnosti prošla tak bouřlivým rozvojem jako shromažďování biologických dat v posledních 30 letech – speciálně to pak platí pro sekvenování DNA molekul.

Velkým impulzem pro rozvoj sekvenačních metod a metod k analýze těchto dat bylo rozhodnutí popsat kompletní genetickou informaci člověka v rámci projektu Human Genome Project. Projekt stál tři miliardy dolarů a byl oficiálně ukončen r. 2000. Rozvoj sekvenačních metod však teprve nabíral na obrátkách – už v r. 2007 sekvenování lidského genomu zlevnilo 3 000× na jeden milion dolarů a další pád

Tab. 1 Přehled základních databází různých typů molekulárních dat

Databáze	Typ dat	Velikost	Webová adresa
ENA	nukleotidová databáze	5 600 000 000 sekvencí	https://www.ebi.ac.uk/ena/browser/home
Uniprot	proteinová databáze	252 000 000 sekvencí	https://www.uniprot.org/
PDBe	databáze 3D struktur	236 000 struktur	https://www.ebi.ac.uk/pdbe/
Ensembl	genomová databáze	36 000 genomů	https://www.ensembl.org/index.html
RNAcentral	databáze nekódujících RNA	35 500 000 sekvencí	https://rnacentral.org/



o tři řády na 1 000 dolarů přišel okolo r. 2015. V dnešní době je sekvenování genomu běžnou službou, kterou lze koupit za 200 dolarů, tedy asi 4 500 Kč. Logickým důsledkem poklesu ceny je, že se sekvenuje víc. Sekvenovaná data je třeba ukládat a to není vzhledem k jejich množství triviální úkol. Jeden lidský genom lze zapsat pomocí přibližně tří miliard písmen – na disku tedy zabere asi tři GB dat. Pokud bychom chtěli uložit genom všech obyvatel České republiky, potřebovali bychom přibližně kapacitu 30 PB dat (PB – petabyte, 10^{15} bytů). Pro srovnání je možné uvést, že Národní knihovna prochází digitalizací a v plánu je zdigitalizovat 310 milionů stran, což představuje asi 1,5 PB dat.

Kompletní genetickou informaci všech obyvatel ČR (zatím) nemáme k dispozici, to však neznamená, že bychom měli málo dat. Velká evropská databáze DNA sekvencí ENA (European Nucleotide Archive, https://www.ebi.ac.uk/ena/browser/home) ukládala v r. 2024 více než 60 PB dat a převyšovala tak mnohonásobně kapacitu naší Národní knihovny. Důležitá je i rychlost, jakou databáze roste – v současnosti se její velikost zdvojnásobí průměrně jednou za 21 měsíců (obr. 5). Databáze ENA však neobsahuje všechny známé sekvence DNA – především genomové sekvence získané v rámci diagnostických vyšetření v nemocnicích nejsou vzhledem k citlivé povaze v této veřejné databázi samozřejmě dostupné. Zjistit množství generovaných sekvenačních dat je proto prakticky nemožné.

Databáze dalších typů molekulárních dat nejsou zdaleka tak velké jako ty nukleotidové, ale rostou někdy díky rozvoji metod překvapivě rychle. Databáze experimentálně určených struktur PDB měla po 25 letech uloženo jen 7 struktur, následně během 50 let nashromáždila kolem 250 tisíc struktur, ale pokrok v predikci 3D struktury (podrobněji na str. 116–118 této Živy) vedl k vývoji databáze predikovaných 3D struktur proteinů, která doslova přes noc v srpnu 2022 zvětšila počet dostupných strukturních dat ze stovek tisíc na bezmála 300 milionů struktur.

5 Rychlost růstu počtu známých sekvencí DNA – množství uložené v databázi ENA v jednotlivých letech. Upraveno podle: European Nucleotide Archive (ENA), kreslila R. Bošková.
6 Infrastruktura Elixir poskytuje biologická data a na webu české pobočky najdete databáze a nástroje vyvinuté v ČR.

Dostupnost biologických dat tedy prochází doslova revolucí – jejich množství roste bezprecedentním tempem. Nové nároky jsou kladeny i na uživatele biologických dat. Pokud byste se rozhodli si celou databázi predikovaných struktur, AlphaFold, stáhnout, počítejte, že stahování potrvá asi 2,5 dne, pokud budete mít velmi rychlé datové připojení (1 GB za sekundu), a zabere spoustu místa na disku, kam data uložíte.

Využití potenciálu dostupných dat je velkou příležitostí, ale i výzvou – je kriticky závislé na kvalitě a dostupnosti výpočetní infrastruktury, na které se data ukládají.

V Evropě jsou data shromažďována především v Evropském bioinformatickém institutu (EBI, https://www.ebi.ac.uk/) v Hinxtonu u Cambridge ve Velké Británii. EBI udržuje nejen ENA a AlphaFold databázi, ale i největší databázi proteinových sekvencí UniProt a více než 50 dalších zdrojů, popisujících široké spektrum biologických dat. EBI má kapacitu asi 400 PB, tedy asi 260 našich Národních knihoven. Ve Spojených státech amerických jsou data ukládána především v rámci National Centre for Biotechnology (NCBI, https://www.ncbi.nlm.nih.gov/), které provozuje 31 databází, v r. 2024 zde bylo kolem 4,6 miliardy nejrůznějších položek. Data ukládaná v Evropě a v Americe se částečně překrývají a velké databáze jako nukleotidová či strukturní jsou vyvíjeny v USA a v Evropě do značné míry nezávisle, ale ukládaná data sdílejí, a tak vzniká i jakási přirozená záloha. Data jsou do určité míry replikována i jinde – např. strukturní data mohou být ukládána ještě v Japonsku a Číně.

EBI i NCBI poskytují data bezúplatně, na rozdíl třeba od řady chemických databází. Prakticky veškerá biologická data (s výjimkou citlivých lidských) jsou k dispozici komukoli. Provoz však něco stojí – zaplatit je třeba datová úložiště, výpočetní kapacity, ale i zaměstnance, kteří se starají o správu dat a další vývoj databází. U proteinové databáze UniProt se odhadují roční náklady okolo 15 milionů eur.

Udržování je náročné ale i po technické stránce – je čím dál těžší shromažďovat všechna data na jednom místě, proto vznikla před 10 lety infrastruktura ELIXIR (https://elixir-europe.org/), jejímž cílem je distribuovat náklady spojené s ukládáním a správou dat mezi co nejvíce členů a zároveň dostat data ještě blíž k uživatelům. Hlavním místem ukládání zůstává EBI, ale na infrastruktuře se podílí v tuto chvíli přes 240 dalších institucí z 21 zemí. Česká republika je jednou ze zakládajících zemí, která zde působí nadále velmi aktivně. ELIXIR CZ (https://www.elixir-czech.cz/, obr. 6) zastupuje 14 institucí a provozuje 16 databází a 43 nástrojů pro jejich analýzu.

Distribuovaná infrastruktura ELIXIRu bude umožňovat sdílení alespoň části citlivých lidských dat, která by jinak kvůli striktní legislativě nemohla opustit hranice státu, ale často ani stěny instituce, v níž byla získána. Příkladem v rámci ELIXIRu je European Genome-Phenome Archive.

Biologie prochází v poslední době doslova revolucí – k dispozici máme zásluhou nových technologií bezprecedentní množství dat, která navíc vyžadují nové dovednosti při analýze, ale i k uchovávání a poskytování dalším členům komunity. Úložiště dat se stala klíčovou součástí vědeckých projektů, a to ve fázi designu experimentů i generování nových dat. Biologické databáze jsou maximálně otevřené a (až na drobné výjimky) zcela zdarma. Často tak slouží i k výuce třeba už na středních školách. Rychle rostoucí množství dostupných dat však vedlo ke změně strategie, kdy se neukládají jen v původních centralizovaných úložištích, ale zodpovědnost je rozložena v rámci rozsáhlé infrastruktury mezi mnoho poskytovatelů, aby byla data dlouhodobě a spolehlivě poskytována široké paletě uživatelů.