4. Informační databáze a jejich využití v experimentální biologii

4.1. Všeobecné databáze sekvencí – nástroje systémové biologie

Kromě bibliografických informací (viz. kapitola 3.) se v moderní biologii do databází ukládají prakticky veškeré dostupné výsledky experimentů prováděných vědci po celé planetě. Tyto informace pak mohou být využívány buď jako široce dostupný zdroj sekvenční informace při studiu námi zvoleného proteinu či proteinového komplexu, mohou však také vzhledem ke své komplexitě sloužit ke studiu regulačních drah či fungování celého organismu. V tom případě je nutné aplikovat též určitý **matematický aparát** a hovoříme o tzv. systémovém přístupu či **systémové biologii**. V zásadě lze databáze hodnotit především podle jejich správy, jak často jsou aktualizovány a hlavně jak účinně filtrují chybné či nesmyslné záznamy.

V případě potřeby získání informace o sekvenci určitého genu či proteinu nejlépe slouží integrovaná databáze **Entrez** (http://www.ncbi.nlm.nih.gov/Entrez/index.html) sdružující hlavní tři **primární databáze GenBank** (USA)**, EMBL-EBI** (Evropa) (http://srs.ebi.ac.uk) **a DDBJ** (Japonsko) (http://www.ddbj.nig.ac.jp). Tato databáze obsahuje celou řadu nástrojů pro hledání literárních a sekvenčních záznamů a také nástroje pro práci se sekcencí samotnou (např. predikce struktur proteinů). Nevýhodou těchto velkých databází je zejména to, že sekvence proteinů jsou zhusta hypotetické, dané pouze predikcí na základě nukleotidových sekvencí, anotace jsou často neúplné či mylné a často je též hůře ošetřena duplicita výsledků.

V případě, že pracujeme na konkrétním experimentálním modelu, vyplatí se na prohledávání sekvencí jeho genomu používat některé více specializované a **spolehlivěji anotované** databáze. Takové jsou stránky **The Institute for Genomic Research (TIGR)**, http://www.tigr.org, či **Sanger Institute** http://www.sanger.ac.uk, zaměřené na **genomiku** obecně. Pro proteomiku a vyhledávání spolehlivých sekvencí proteinů je výhodné pracovat na dobře spravované databázi v rámci http://www.expasy.org/.

Pro práci se sekvenčními daty lze v zásadě využívat jakýkoliv program typu **Notepad**. Je ovšem pravda, že lze na webu nají velké množství specializovaných nástrojů jako je např. http://workbench.sdsc.edu/, kde lze data třídit, porovnávat, měnit jejich formu atd. Poměrně dobrým nástrojem je i projekt pod **EBI**, tzv. **Biomart Project** (http://www.biomart.org/index.html), kde lze poměrně jednoduchým způsobem hromadně sekvence hledat, zpracovat a exportovat. Existují i komerční programy, kde lze sekvenční data porovnávat, upravovat, navrhovat primery, konstruovat mapy atd.

4.2. Některé specializované sekvenční databáze

V případě, že nás již zajímá konkrétní organismus a chceme se zaměřit pouze na hledání v jeho sekvenčních údajích či v údajích z příbuzných organismů, je lépe využít specializované databáze. Ty mají oproti výše uvedeným databázím výhodu v tom, že bývají zpravidla daleko lépe spravovány a případné chyby jsou daleko účinněji odstraňovány. Příkladem může být databáze zaměřená na oblíbený model rostlinu *Arabidopsis thaliana* na adrese http://www.arabidopsis.org - **The Arabidopsis Information Resource (TAIR)** nebo databáze čeledi Solanaceae http://www.sgn.cornell.edu/.
Informace o přepisu konkrétního genu tj. informace o tzv. **transkriptomu** lze v současné době u rostlin nejlépe získat na stránkách projektu **Genevestigator**, https://www.genevestigator.ethz.ch/. Tyto stránky sdružují informace z experimentů prováděných s využitím **čipové technologie** (sledující expresní profil obrovského množství genů najednou) a představují neocenitelný zdroj informací. Na stránkách genevestigatoru jsou v současné době pouze informace z myši a *Arabidopsis*, ostatní organismy je potřeba

hledat s pomocí obecných databází či např. přes Google (klíč. slova jako yeast transcriptome, atd..). Příkladem pěkně spravované databáze expresních profilů je **Arabidopsis gene family profiler,** http://aarabidopsisgfp.ueb.cas.cz/index.php.
Podobně jako pro transkriptomická data, existují i pro **proteomická** data specializované databáze. Často se lze setkat s přístupem třídit a anotovat dostupné informace podle charakteru studovaného proteinu. Takto existují databáze rostlinných membránových proteinů (http://aramemnon.botanik.uni-koeln.de/) či ještě specializovanější http://www.cbs.umn.edu/arabidopsis/. Omezit se lze i na konkrétní organelu na její proteom prohledávat, např rostlinné plastidy na stránkách http://www.plprot.ethz.ch/.
Velmi zajímavým přístupem je organizovat informaci o proteomu v integrované databázi, která kromě sekvenční informace obsahuje i informaci o **lokalizaci konkrétního proteinu** v buňce. Tyto databáze jsou teprve ve stádiu zrodu, ale dá se očekávat, že v budoucnu budou takové informace běžně dostupné např. přes pubmed. Příkladem databáze lokalizací u rostlin je http://aztec.stanford.edu/gfp/index.html, kde lokalizační informace pocházejí z *in vivo* pozorování tagovaných proteinů.
Při rutinní laboratorní práci se též velmi hodí databáze zpracovávající informace o tom, jaký je projev mutace urč. proteinu, tzv. databáze knockoutů (**knockout databases**). Mutanty připravené pomocí technologie RNAi lze např. hledat na http://www.agrikola.org/index.php?o=/agrikola/main.