Pokroky matematiky, fyziky a astronomie

Fatima Cvrčková

Jak se čtou genomy: bioinformatika jakožto obor na pomezí biologie a exaktních věd

Pokroky matematiky, fyziky a astronomie, Vol. 51 (2006), No. 4, 288--300

Persistent URL: http://dml.cz/dmlcz/141329

Jak se čtou genomy: bioinformatika jakožto obor na pomezí biologie a exaktních věd

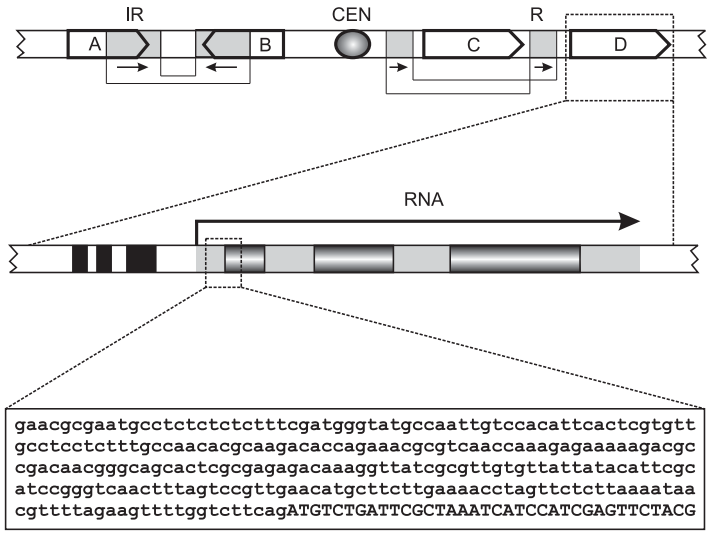
Fatima Cvrčková, Praha

1. O čem je bioinformatika

V roce 2006 uplynulo již čtyřicet let od rozluštění genetického kódu, tedy vztahu mezi pořadím bází nukleotidových monomerů („písmenek“) v řetězci DNA a pořadím aminokyselin v molekule bílkoviny, která je příslušným úsekem DNA kódována ([19], [22]; viz též [18]). Znalost kódu pochopitelně přinesla motivaci pro vývoj postupů umožňujících experimentálně zjišťovat pořadí (sekvenci) bází v DNA, a to pokud možno ve velkém — tak, abychom mohli „přečíst“ i kompletní sekvenci „genetického receptu“ (genomu) zkoumaného organismu. První genom živé buňky, totiž bakterie *Haemophilus influenzae*, byl přečten před více než 10 lety [12] a postupem doby přibyla řada dalších, mikrobiálních, rostlinných i živočišných (včetně genomů několika desítek druhů bakterií, pivní kvasinky [15], huseníčku Thalova [26], člověka [27], [28], myši [21], rýže [14], [29], a nedávno i šimpanze [6]). Právem tedy lze dnes hovořit o tom, že žijeme v postgenomové době.

Naléhavou se ovšem stává otázka po smyslu přečteného. Nemá-li být zjištěná sekvence bází pouze pomníkem naší technické dovednosti (přiznejme, že impozantním: 140 milionům párů bází genomu skromného plevele huseníčku by odpovídalo zhruba 78 000 standardních strojopisných stran pokrytých výhradně znaky A, C, G a T zastupujícími báze adenin, cytosin, guanin a thymin), musíme se v ní umět nějak vyznat a pochopit, co která část předlouhého řetězce znaků znamená. Na sekvenci genomu se můžeme dívat jako na text napsaný sice známou abecedou, ale v neznámém jazyce. Víme, jak přepisovat jednotlivá písmena do latinky, občas rozluštíme slovo a místy chápeme i zákonitosti toho, jak se slova skládají do vět, ale obsah celého textu sotva začínáme tušit. Rozluštění genetického kódu znamenalo „pouze“ odhalení pravidel, podle nichž buňka v procesu translace („překladu“) odvozuje posloupnost aminokyselin v proteinu z posloupnosti bází DNA postupem, který by bylo možno formálně popsat jako mnohoznačné zobrazení. Kód sám však nic nevypovídá o tom, které úseky DNA budou sloužit jako zápis pořadí aminokyselin a které budou hrát třeba regulační nebo strukturní úlohu, a už vůbec ne o tom, jaká bude biologická funkce

Doc. RNDr. Fatima Cvrčková, Dr. rer. nat. (1966), katedra fyziologie rostlin Přírodovědecké fakulty UK v Praze, Viničná 5, 128 44 Praha 2, e-mail: fatima@natur.cuni.cz



Obr. 1. Příklady motivů, které lze bioinformatickými metodami vyhledávat v sekvencích DNA. Nahoře: schéma části chromozómu s centromerou (specializovanou oblastí, která zajišťuje řízenou distribuci chromozómů při buněčném dělení, CEN) a čtyřmi geny (A, B, C, D). Opakované sekvence — přímé opakování (repetice, R) a obrácené opakování (invertovaná repetice, IR) jsou zvýrazněny, šipky udávají směr čtení opakované sekvence. Nález takových opakování je důležitým zdrojem informací pro rekonstrukci evoluční historie genomů. Uprostřed: detail genu D. Černě regulační oblasti, šipka (RNA) označuje úsek přepisovaný do RNA, která je však posléze buňkou upravována, takže nakonec se do proteinu přeloží pouze stínovaná část. Dole: sekvence odpovídající oblasti v rámečku na prostředním obrázku, úsek překládaný do proteinu velkými písmeny.

kódovaných proteinů (obr. 1). Z experimentálních výsledků získaných během čtyř desetiletí, která od rozluštění kódu uplynula, však bylo již možno vyvodit zákonitosti, které dovolují odhadovat funkci jednotlivých úseků genomu nebo sekvenčních motivů v proteinech.[1])

Právě teoretické zkoumání sekvencí biologických makromolekul je jedním z ústředních témat **bioinformatiky** — oboru na pomezí biologie a informatiky, který se zabývá především zpracováváním, prohledáváním a analýzou dat o sekvenci a struktuře biologických makromolekul, zejména nukleových kyselin (DNA a RNA) a proteinů, samozřejmě za pomoci výpočetní techniky [8].[2]) Typickými úlohami bioinformatiky

[1]) Zájemce o podrobnější výklad biologických souvislostí odkazuji na kteroukoli moderní učebnici molekulární biologie; k analogii sekvence DNA a textu viz též [20].

[2]) Takto chápaná bioinformatika se někdy označuje jako „klasická“. Někteří autoři však bioinformatiku vymezují jako jakékoli „využití počítačů k hledání odpovědí na biologické otázky“ [3]. To by ale zahrnovalo i statistické zpracování fyziologických, klinických nebo dokonce ekologických dat, tedy oblast, která je od studia sekvencí a struktur makromolekul velmi vzdálená jak obsahově, tak i metodicky. Přesně formulovaná konsensuální definice bioinformatiky snad ani neexistuje, za „bioinformatiky“ se s odkazem na „definiční třídění ruských vědců“ považují dokonce i čeští parapsychologové [23]. Dlužno ovšem poznamenat, že odborná bioinformatická komunita od nich udržuje pochopitelný odstup.

jsou vyhledávání funkčně zajímavých sekvenčních motivů (třeba genů kódujících proteiny) v celých genomech a jiných velkých souborech sekvencí, katalogizace sekvencí genů a proteinů podle biologicky významných kritérií (např. evoluční příbuznosti nebo biologické role), vývoj algoritmů a programů pro automatizované zpracování a analýzu sekvenčních dat a srovnávací analýzy celých genomů či vybraných skupin vzájemně příbuzných genů s cílem rekonstruovat evoluční historii (fylogenezi) zdrojových organismů nebo genových funkcí (viz např. [4], [9], [11]).

Interpretace genomových sekvencí je bez bioinformatických metod nemyslitelná. Proto můžeme bioinformatiku považovat za podstatnou součást **genomiky** — odvětví molekulární biologie, které se věnuje analýze genomů. S trochou nadsázky bychom mohli spolu s jedním z nejvýznamnějších molekulárních biologů posledních desetiletí Sydneym Brennerem [5] konstatovat, že jedním z cílů genomiky je nalézt třeba ten příslovečný gen, který může za to, že my mluvíme, kdežto šimpanzi ne[3]). Realističtější představu o obsahu genomického bádání ovšem poskytne nahlédnutí do kterékoli z výše citovaných genomových publikací, jejichž jádrem je obvykle komentovaný katalog genů a sekvenčních motivů nalezených v sekvenovaném genomu.

Bioinformatika se však nezabývá pouze sekvencemi. Dnes jsou již k dispozici metody, které dovolují souběžně sledovat aktivitu až desítek tisíc různých genů stanovením hladin příslušných RNA transkriptů. Analýza záplavy dat, která takovéto experimenty produkují, se opět neobejde bez složitých výpočetních postupů. Příslušná oblast molekulární biologie, běžně označovaná jako **transkriptomika**, si tudíž vyžádala vznik specifického odvětví bioinformatiky. Podobně i **proteomika**, obor, který se zabývá studiem proteinového složení buněk a jeho proměn v nejrůznějších biologicky zajímavých situacích, jako je třeba embryogeneze nebo nádorové zvrhnutí (transformace), se nemůže obejít bez bioinformatických metod analýzy dat.

Zejména v poslední době nabývá na významu také teoretické studium prostorových struktur biologických molekul, především proteinů (tzv. **strukturní bioinformatika**). Tato oblast se stala ohniskem zájmu farmakologů, kteří využívají znalostí o struktuře buněčných regulačních proteinů k návrhu specifických léčiv, a lze ji v jistém ohledu již považovat za speciální odvětví teoretické chemie.

[3]) „*... Kdysi jsem si představoval, že bychom měli sekvenovat oba genomy — lidský i šimpanzí — a odečíst je od sebe. Našli bychom tak gen pro řeč, protože tato schopnost nás od šimpanze odlišuje dost významně. Ale nebylo by tak úplně jasné, u kterého druhu bychom ten gen našli. Může chybět u šimpanzů a být přítomen v nás — pak by se takový gen specifický pro člověka mohl nazývat »Chomského genem«. Anebo naopak bychom ho postrádali my, protože co když šimpanzi nahlédli, že lepší bude být zticha? Pak bychom ho nazvali »Chimpského gen«* ... ([5], překlad A. Markoš). S. Brenner je nositelem Nobelovy ceny za biologii a lékařství z r. 2002 za významný příspěvek k pochopení vývoje (ontogeneze) mnohobuněčných organismů, zejména za zavedení a charakterizaci hlístice *Caenorhabditis elegans* jakožto modelového organismu.

2. Biologie není matematika

Na bioinformatiku můžeme právem pohlížet jako na odvětví, v němž se stýkají dva světy se zásadně odlišnou tradicí: svět exaktních oborů — matematiky a informatiky — a svět praktického přírodozpytu. Zvyk členit vědní obory po angloamerickém vzoru na „přírodovědné“ (*sciences*) a „humanitní“ (*humanities*), přičemž matematika jaksi samozřejmě zapadá do přírodovědné krabičky, nás může svádět k tomu, abychom význam propastného rozdílu mezi matematickým a biologickým pohledem podceňovali. Připomeňme si však, že uvedené členění není jediné možné — tak třeba na univerzitě ve Vídni se obory, které jsme si zvykli chápat jako přírodovědné, pěstují na fakultě „formálních a přírodních věd“ (*Formal- und Naturwissenschaftliche Fakultät*).

Setkávání rozdílných oblastí vědy a s ním spojená nezbytnost komunikace odvětví, která se po staletí vyvíjela odděleně, a tudíž si každé z nich vybudovalo svůj vlastní jazyk, tradici a kulturu toho, co se v odborném textu sluší a co ne, ovšem není prosté úskalí. Podle významného biologa a popularizátora matematiky 1. poloviny 20. století Lancelota Hogbena hovoří matematika „řečí veličinnou“ či „řečí popisující míru věcí“ na rozdíl od „obyčejných řečí, jimiž popisujeme druh věcí ve světě“ [16]. Biolog, zvyklý zabývat se právě pozorováním druhu věcí ve světě, článku napsanému matematikem často nerozumí z důvodů v podstatě jazykových, a někdy ani nepochopí, proč by ho téma vůbec mělo zajímat, a asi i naopak. Je proto snad na místě pokusit se — ovšem v duchu biologické, nikoli matematické tradice! — vymezit nejnápadnější rysy, které již na první pohled odlišují obvyklý styl výkladu v jazyce matematickém (tak, jak jsou na něj čtenáři tohoto časopisu zvyklí), od způsobu běžného v biologii.

První, co pravděpodobně nematematika při srovnání biologického a matematického textu upoutá, je absence „tvrdých“ definic a formalizované struktury výkladu v biologii. V biologických oborech prakticky není vět a důkazů a snad veškerá tvrzení, závěry, vývody, zákonitosti, a dokonce i definice provází nějaké to „obyčejně“, „většinou“ či „zpravidla“. Biologický výklad je (odhlédneme-li od specifického slovníku, dnes často bohatého na zkratky upomínající prý na vojenskou terminologii) veden měkkým jazykem blízkým každodenní řeči.

Biologie dospívá k závěrům zobecňováním dílčích pozorování, tedy induktivně. Extrapolace se všemi nebezpečími, která z ní plynou, je přijímána jako veskrze legitimní postup. Biolog se však ne vždy stará zvláště o vymezení oboru platnosti svých tvrzení a pečlivé určení jeho hranic. Meze mohou být dokonce mapovány až *ex post*, ve chvíli, kdy pojmeme podezření, že jsme se jim nebezpečně přiblížili nebo je překročili. Tak například řada našich současných představ o pochodech, které pohánějí evoluci, vzešla ze zobecnění zákonitostí odvozených ze studia pohlavně se rozmnožujících živočichů (viz např. [30]). Úvahy a diskuse o tom, nakolik se od zavedeného schématu liší jiné organismy, například mikrobi nebo nepohlavně se množící rostliny, navazují až posléze, tak, jak jsou postupně nalézány další a další důvody možných odchylek.

Biologie, včetně „biologické“ větve bioinformatiky, navíc zaujímá vůči matematice postoj, který by se snad dal označit jako inženýrský, nebo možná dokonce i kutilský. Není to přitom míněno nijak urážlivě: metafora „kutění“ (*bricolage*, *tinkering*) hraje v molekulárně biologickém uvažování dosti významnou úlohu. Věříme totiž, že nové

biologické funkce (a molekulární aparát, který je zajišťuje) vznikají jako výsledek kombinování a modifikací stávajícího „vybavení“ organismu metodou pokusu a omylu. Bakterie experimentuje se svými geny podobně jako kutil s obsahem bezedných přihrádek ve své dílně, programátor s již fungujícími procedurami a moduly nebo bioinformatik s arzenálem dostupných výpočetních postupů, dokud se nedopracuje něčeho použitelného, byť i ne nutně elegantního ([17], viz též [7]).

Pokud tedy matematika biologa vůbec nějak oslovuje (což v kontextu genomiky a bioinformatiky musíme předpokládat), chápe ji především jako **nástroj**, který může při vhodném používání napomoci v interpretaci údajů získaných pozorováním či experimentálním studiem hmotného světa. Biologa neupoutá třeba možnost vyjádření genetického kódu pomocí monoidů [18], pokud není zřejmé, na jaké konkrétní *biologické* otázky takový popis může pomoci odpovědět.[4]) Vztah biologa k matematice je tedy podobný pohledu technika. Není důležité, jaký teoretický postup byl použit k výpočtu potřebné tloušťky výztuže mostního oblouku, a nezáleží ani na tom, zda výpočet byl exaktní nebo přibližný, elegantní nebo upachtěný. Důležité je pouze dvojí: aby most bylo možno prakticky postavit a aby nespadl. Pokusím se to ilustrovat na příkladu toho, jak bioinformatik uvažuje, když řeší poměrně obvyklý problém vyhledávání vzájemně příbuzných sekvencí. To, že nadále diskutované postupy možná (aspoň pro někoho) nejsou prosté určitého druhu krásy, je sice příjemné, avšak zdaleka ne zákonité.

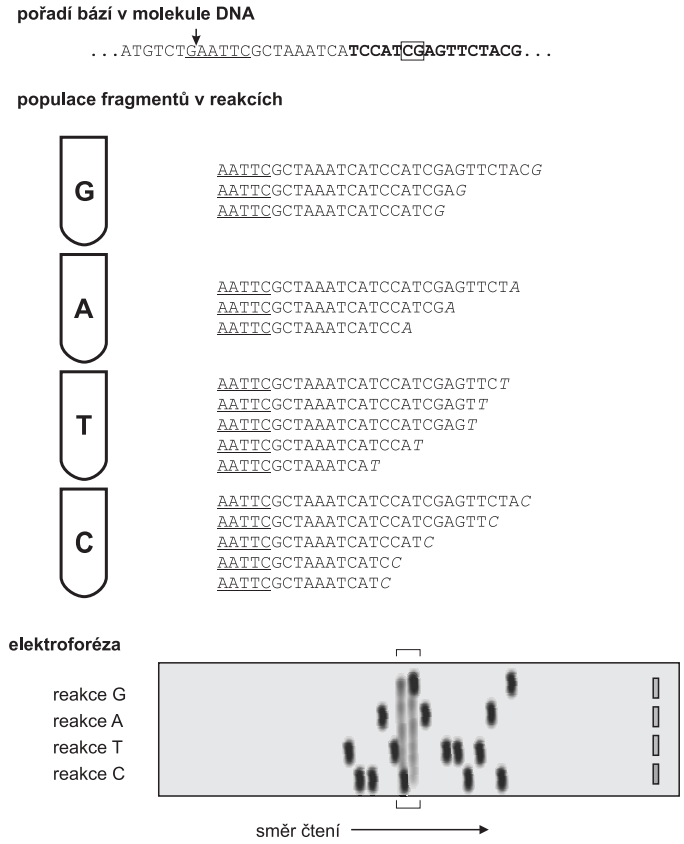
3. Digitální a tělesné

Dosud probírané rozdíly v přístupu biologa a matematika k předmětu jeho bádání se sice mohou jevit jako podstatné, avšak jsou pouhými důsledky rozdílu jiného, zcela zásadního. Zatímco biologie se zabývá hmotnými (a dokonce živými) objekty z našeho **tělesného světa**, pro jejichž chování můžeme pouze více či méně úspěšně vyvozovat pravidla na základě pozorování a experimentů, matematika působí ve světě (či spíše zásvětí) **virtuálních myšlenkových konstruktů**, pro něž lze zákonitosti stanovit předem.[5]) Matematické modelování i jiné využití matematiky v biologii závisí na možnosti nalézt takové konstrukty, které se v nějaké důležité vlastnosti shodují se zkoumaným objektem, a přitom se s nimi současně ve virtuálním zásvětí dobře pracuje.

V případě té části bioinformatiky, která se zabývá „čtením“ genomů, pak vhodným modelem jsou *digitální sekvence* znaků z předem definované množiny, která v případě

[4]) Stojí zde za to citovat *in extenso* výstižný vhled již zmiňovaného L. Hogbena, jednoho z průkopníků aplikace matematických a statistických metod v genetice: *„Na každém dějinném stupni vědy musíme rozlišovat dvojí stanovisko, kterému se říká vědecké. Jedno zahrnuje soud o řádu přírodním založený na pozorování. Druhé se skládá z matematických představ nebo obrazů podnícených početními návody nebo použitých k tomu, aby podnítily početní návody, kterých se používá k práci s nimi. První jsou trvalými kameny vědecké stavby. Druhé jsou dočasným lešením, kterého se použije a které se strhne podle toho, jak stavba pokračuje.“* ([16], s. 523)

[5]) Zájemce o hlubší rozbor této problematiky odkazuji na práce Z. Neubauera, P. Vopěnky a Z. Kratochvíla.



Obr. 2. Postup, kterým lze od tělesné molekuly DNA dospět k digitální sekvenci znaků, ilustruje schéma klasické Sangerovy metody sekvenování DNA, od níž je odvozena většina dnes používaných technik. Prvním krokem je vytvoření čtyř souborů definovaných krátkých fragmentů DNA v rozmezí od několika málo desítek po několik stovek bází, které sdílejí společný začátek. Tímto začátkem může být například oblast DNA rozpoznávaná specifickým bakteriálním enzymem (třeba restrikční endonukleázou, která štěpí trojrozměrnou strukturu molekuly odpovídající podtržené sekvenci v místě šipky). Ve čtyřech chemických reakcích katalyzovaných izolovanými enzymy bakteriálního původu vznikají fragmenty, které vždy končí v pozici odpovídající určité bázi (kurzívou). Následuje elektroforetické dělení populací molekul podle velikosti s přesností, která umožňuje odlišit fragmenty, které se délkou liší o jediný nukleotid (větší fragmenty putují pomaleji, na obrázku tedy elektroforéza postupuje zprava doleva). Z polohy skvrn po vizualizaci fragmentů lze pak přímo odečíst sekvenci (na obrázku je čtený úsek tučně). Chování fragmentů při elektroforéze však odráží nejen jejich velikost, ale i tvar (hlavně sklon k vytváření stabilních sekundárních struktur) a velké populace molekul se chovají do jisté míry nahodile. Proto je výsledný obraz nutně mírně rozmytý a v principu nikoli digitální, ale analogový. Jeho digitalizace nemusí být vždy triviálním úkolem (viz oblast elektroforetogramu vyznačená hranatou závorkou, odpovídající sekvenci v rámečku).

sekvencí DNA zahrnuje minimálně čtyři symboly (A, C, G, T) zastupující čtyři báze, v případě proteinů pak minimálně 20 symbolů pro dvacet standardních aminokyselin. Někdy ale pracujeme i se širším souborem znaků, aby bylo možno zapisovat sekvence

s nejednoznačnými pozicemi (tak např. Y znamená podle standardizované tabulky Mezinárodní unie pro čistou a aplikovanou chemii IUPAC „buď C, nebo T“, N znamená libovolnou bázi). Důvodem nejednoznačnosti může být buďto potřeba zaznamenat současně sekvenci několika variant zkoumané molekuly, nebo to, že se nám experimentálně nepodařilo jednoznačně zjistit pořadí bází ve studovaném úseku sekvence. Protože genomika pracuje nejčastěji s molekulami DNA, zaměříme se v následujícím výkladu právě na tento případ; mnohé však platí obdobně i pro RNA, případně proteiny.

Jakým způsobem vůbec můžeme převést strukturu složité molekuly na sekvenci digitálních znaků? Na molekulu DNA lze pohlížet jako na nosič, na kterém je zapsána digitální, „nehmotná“ sekvence [30]. Současně však DNA je hmotnou, tělesnou molekulou, jejíž praktické zkoumání, včetně „čtení“ sekvence, je založeno na postupech mokré (a leckdy nelibovonné) chemie a biochemie.[6]) Interpretace výsledků analýzy, jejíž součástí je i netriviální digitalizace analogového výstupu, nemusí být vždy zcela přímočará (obr. 2). V ideálním případě bychom sice měli získat autentickou sekvenci, avšak ve skutečnosti se vždy musíme spokojit s nejlepší možnou, i když ne nutně zcela správnou variantou čtení analogového záznamu.

Jedním z centrálních postupů praktické bioinformatiky je porovnávání dvou či více sekvencí a zjišťování míry jejich vzájemné **podobnosti**. Zastavme se proto u otázky, jak vůbec podobnost chápeme a jak můžeme stanovovat kvantitativní míru podobnosti dvou sekvencí.

Samo vymezení toho, co vlastně znamená podobnost ve světě digitálních sekvencí, přitom není samozřejmé. Snadno nahlédneme, že pokud nepřijmeme nějaké další předpoklady, mohou dva řetězce digitálních znaků být buď totožné, nebo nestejné: shodují-li se ve všech znacích, jsou totožné, liší-li se byť i jen v jediném znaku, už jsou to různé řetězce.

V bioinformatickém kontextu však digitální znaky vystupují jakožto **symboly**, zastupující tělesné objekty, například nukleotidy či aminokyseliny nebo jejich řetězce. Proto má smysl hovořit zde i o podobnosti v tom smyslu, jak ji chápeme v běžném životě (viz [10]). V tělesném světě totiž typicky nalézáme (a dobře rozpoznáváme) nikoli vztahy totožnosti (identity) a nestejnosti, ale víceméně odstupňovanou „podobnost“. I takzvaně identická (jednovaječná) dvojčata se sobě navzájem pouze podobají a *nejsou* totožná. Co je pro posuzování podobnosti relevantní, přitom musí rozhodnout rozumějící (a rovněž tělesný) pozorovatel. Podobnost má *vždy* smysl pouze v konkrétním kontextu, který určuje, na čem záleží a na čem ne.[7])

[6]) Nesmíme také zapomenout, že vzorek, který zkoumáme, a dokonce i mnohé reagencie v experimentech používané jsou odvozeny z živých organismů (tak například jako velmi specifická analytická činidla se v molekulární biologii využívají enzymy, zpravidla získávané z bakterií — viz obr. 2 a 3). I laik si asi dovede představit, jak vážná bývá například hrozba mikrobiální zkázy reagencií.

[7]) V jistém smyslu to platí i pro zdánlivě bezkontextovou podobnost geometrickou (např. podobnost trojúhelníků). Tam totiž kontextem je sama geometrie — můžeme si představit i situaci, kdy trojúhelníky v tradičním geometrickém smyslu nepodobné vystupují jako podobné (např. máme-li vybrat „vzájemně podobné obrazce“ z množiny, která obsahuje trojúhelníky a čtyřúhelníky).

V našem konkrétním případě, kde digitální znaky zastupují jednotlivé báze v řetězci DNA či aminokyseliny v molekule proteinu, můžeme pro libovolnou dvojici sekvencí vzájemně shodné délky určit míru podobnosti následujícím postupem:

- Sekvence přiložíme k sobě po celé délce, to znamená zapíšeme je do dvou pod sebou umístěných řádků tak, aby sobě odpovídající pozice ležely pod sebou a vytvořily tedy dvojici (takový zápis označujeme jako přiřazení — *alignment*).
- Každé dvojici znaků, která se v přiřazení může vyskytnout, přiřadíme konkrétní číselnou hodnotu, která vyjadřuje vzájemnou podobnost aminokyselin či nukleotidů, které tyto znaky zastupují. V nejjednodušším případě můžeme jakékoli identické dvojici pozic (páru) přidělit hodnotu 1 a jakékoli neidentické dvojici (nepáru) hodnotu 0.
- Celkovou hodnotu podobnosti stanovíme jako součet hodnot podobnosti všech jednotlivých pozic přiřazení. Chceme-li vzájemně porovnávat míru podobnosti různě dlouhých dvojic sekvencí, normalizujeme takto zjištěnou hodnotu podobnosti vydělením počtem pozic v přiřazení.

Uvedený základní postup přiřazuje identickým sekvencím maximální podobnost (při uvedených hodnotách parametrů by normalizovaná hodnota podobnosti byla 1), avšak i dvěma náhodně vybraným náhodným sekvencím odpovídá nenulová hodnota podobnosti (v takovém případě bychom např. pro DNA při rovnocenném zastoupení všech bází očekávali průměrnou normalizovanou podobnost 0,25). Z toho plyne, že bude nutné stanovit rovněž kritéria pro to, jakou míru podobnosti lze ještě pokládat za biologicky významnou. Tento postup lze též pokládat za východisko pro složitější a obecnější varianty, použitelné i pro reálné situace, kde např. většinou pracujeme se sekvencemi různé délky:

- Liší-li se porovnávané sekvence délkou, zavádíme do jedné či obou sekvencí mezery tak, abychom pokud možno maximalizovali výslednou míru podobnosti. Přitom je rozumné předpokládat, že sám akt zavedení mezery i její prodlužování míru podobnosti snižuje; pro výpočet tohoto snížení („ceny za mezeru“) zavádíme empirické parametry.
- Vzhledem k tělesné povaze aminokyselin a nukleotidů nemusíme považovat všechny nepáry za rovnocenné. Dvojicím, které jsou si vzájemně podobné tvarem, velikostí a nábojem molekuly, můžeme přisoudit vyšší („podobnější“) míru podobnosti než dvojicím diametrálně odlišným. Právě tak můžeme brát v úvahu i frekvenci výskytu jednotlivých aminokyselin v empiricky zjištěných sekvencích a přiřazovat shodě vzácných znaků větší váhu než těm hojným. Obojí lze vyjádřit prostřednictvím substituční matice — tedy tabulky, která každé možné dvojici symbolů přiřazuje konkrétní číselnou hodnotu podobnosti.

Ani cenu za mezeru, ani substituční matici nelze odvodit teoreticky z nějakých výchozích předpokladů. Tyto parametry je nutno odvodit empiricky na základě analýzy reálných sekvencí, které se liší jen velmi málo. Předpokládáme totiž, že takové velmi

podobné sekvence prakticky mohly vzniknout jen poměrně nedávnou evoluční divergencí ze společného předka, a proto je jejich podobnost zřejmě biologicky významná.[8])

4. Jak se neutopit v datech

Vzájemné porovnávání sekvencí a stanovování míry jejich vzájemné podobnosti je jádrem celé řady úloh, kterými se bioinformatikové zabývají. Prakticky každý, kdo někdy sekvenoval kousek DNA, se totiž zajímá o to, čemu již známému se jeho nová sekvence podobá, to jest zda má nějaké příbuzné ve veřejně dostupných databázích. Od takového zjištění se totiž pak odvíjí plánování dalších experimentů, otvírá se cesta k porovnávání vlastností zkoumaného úseku s geny známé funkce a podobně. (Také identifikaci opakovaných úseků v rámci genomů lze v zásadě převést na vyhledávání na základě podobnosti.)

Základy bioinformatiky byly naštěstí položeny v době, kdy měřítkem kvality výzkumu byly publikace spíš než patenty a volné sdílení primárních dat se stalo samozřejmostí, od níž snad již nelze ustoupit. Lepší odborné časopisy dodnes vyžadují, aby autoři, kteří publikují nové sekvence, tato data vložili do některé z veřejných databází. Nukleotidové sekvence spravuje a zpřístupňuje zejména konzorcium spojující americkou databázi GenBank, evropskou EMBL (European Molecular Biology Laboratory Data Library) a japonskou DDBJ (DNA Data Bank of Japan). Tyto tři databáze udržují společný standard identifikace jednotlivých záznamů, průběžně si vyměňují data a jejich obsah je až na přírůstky z posledních několika hodin totožný. V současnosti obsahuje „velká trojka“ databází konzorcia desítky milionů záznamů úhrnné délky zhruba sta miliard bází, a dat neustále přibývá.[9])

Už sama údržba tak velkých databází je úctyhodným úkolem, protože software i hardware musí stále držet krok s dramatickým růstem objemu dat, a o **vyhledávání** v nich to platí v míře ještě větší. Podle výstižného příměru z přednášky amerického ekologa Barryho Rocka, který se sice nezabývá bioinformatikou, avšak s podobně velkými databázemi má bohaté zkušenosti, jde o úkol srovnatelný s pitím z vysokotlaké požární hadice („*drinking from the firehose*“).

Ilustrovat to můžeme příkladem běžné úlohy: jak v databázi najít všechny sekvence významně podobné námi zadané sekvenci, kterou budeme dále označovat jako dotaz (*query*). Mohli bychom postupovat třeba takto:

- Pro každou sekvenci z databáze sestrojíme přiřazení vůči dotazu. Protože sekvence z databáze a dotaz se v obecném případě liší délkou, musíme přiřazení optimalizovat

[8]) I v rámci standardního myšlenkového rámce moderní molekulární biologie si lze představit alternativní vysvětlení, totiž evoluční konvergenci sekvencí odlišného původu směrem k vzájemné podobnosti např. působením selekčního tlaku. Vznik rozsáhlejších oblastí podobnosti touto cestou je však krajně nepravděpodobný.

[9]) Aktuální stav lze nalézt buď na webových stránkách zmíněných databází (např. `http://www.ncbi.nlm.nih.gov` pro GenBank), nebo v každoročním zvláštním „databázovém“ čísle časopisu Nucleic Acids Research.

a zvolit to, které má co nejvyšší celkovou (nenormalizovanou) hodnotu podobnosti s dotazem; tuto hodnotu vždy zaznamenáme.

- Sekvence z databáze seřadíme podle hodnot podobnosti s dotazem. (Už víme, která sekvence v databázi je dotazu nejbližší, avšak nevíme, zda je dost blízká na to, aby šlo o nenáhodnou podobnost).
- Sestrojíme histogram rozdělení hodnot podobnosti dotazu s jednotlivými sekvencemi z databáze a zjistíme, pro které sekvence z databáze leží hodnoty podobnosti výrazně mimo křivku nahodilé distribuce. Pro tyto sekvence je míra shody s dotazem zjevně větší než nahodilá.

Uvedený značně zjednodušený algoritmus má jednu zásadní nevýhodu: prohledává sice databázi způsobem skutečně vyčerpávajícím, avšak zejména první krok — sestavování a optimalizace přiřazení pro všechny sekvence v databázi — je mimořádně náročný na čas i výpočetní kapacity. Proto se v praxi takovéto „přesné“ algoritmy používají jen zřídka, a častěji se setkáváme s postupy **heuristickými**, které prohledávání výrazně zrychlují, byť i za cenu možné újmy na přesnosti.

V našem příkladu byla jedním z nejpomalejších kroků konstrukce a optimalizace přiřazení mezi dotazem a databázovými sekvencemi. Přitom pro mnohé sekvence by se tento krok dal vynechat, kdybychom uměli míru podobnosti s dotazem zhruba odhadnout a hned vyloučit ty sekvence, které pravděpodobně ani po důkladném ladění přiřazení nedosáhnou vysokých hodnot podobnosti. Jen pro zbývající „kandidáty“ pak přiřazení vypracujeme skutečně pečlivě. Na podobném principu jsou skutečně založeny v současnosti běžně používané algoritmy pro prohledávání velkých databází, na nichž jsou založeny programy, jako je FASTA [25], zrychlující prohledávání databází oproti přesným algoritmům o více než 2 řády, nebo novější a ještě rychlejší BLAST. Princip algoritmu používaného programem BLAST (*Basic Local Alignment Search Tool*) si můžeme představit přibližně takto (viz [1], [2], [13], [24]):

- Program nejprve vytvoří „slovník“ výskytu všech přítomných kombinací symbolů („slov“) předem zvolené délky v sekvenci dotazu (obvykle se používá empiricky osvědčená délka 2–3 aminokyseliny nebo 10–11 nukleotidů). Pro každé slovo (*word*) ze slovníku pak za použití zvolené substituční matice zjistí hodnotu podobnosti se sebou samým a se všemi jeho deriváty vzniklými záměnou jediného symbolu. Ze slovníku pak vypustí všechna slova, jejichž hodnota podobnosti je nižší než předem zvolená mezní hodnota T (*threshold*).
- V dalším kroku program pro každou sekvenci v databázi stanoví výskyt slov ze slovníku, a dále se zabývá pouze těmi sekvencemi, které obsahují minimálně dvě slovníková slova (*hits*) oddělená vzdáleností menší či rovnou, než je předem stanovená „délka okna“ A.
- Každá takto nalezená dvojice sousedících slovníkových slov poslouží jako „jádro“ pro konstrukci přiřazení, které program rozšiřuje oběma směry tak dlouho, dokud hodnota podobnosti (*score*) nezačne klesat. Pokud je takto získaná hodnota podobnosti nižší než předem zvolená mezní hodnota (*cutoff*), program přiřazení opustí; v opačném případě je zaznamenána jako tzv. HSP (*high scoring pair*).

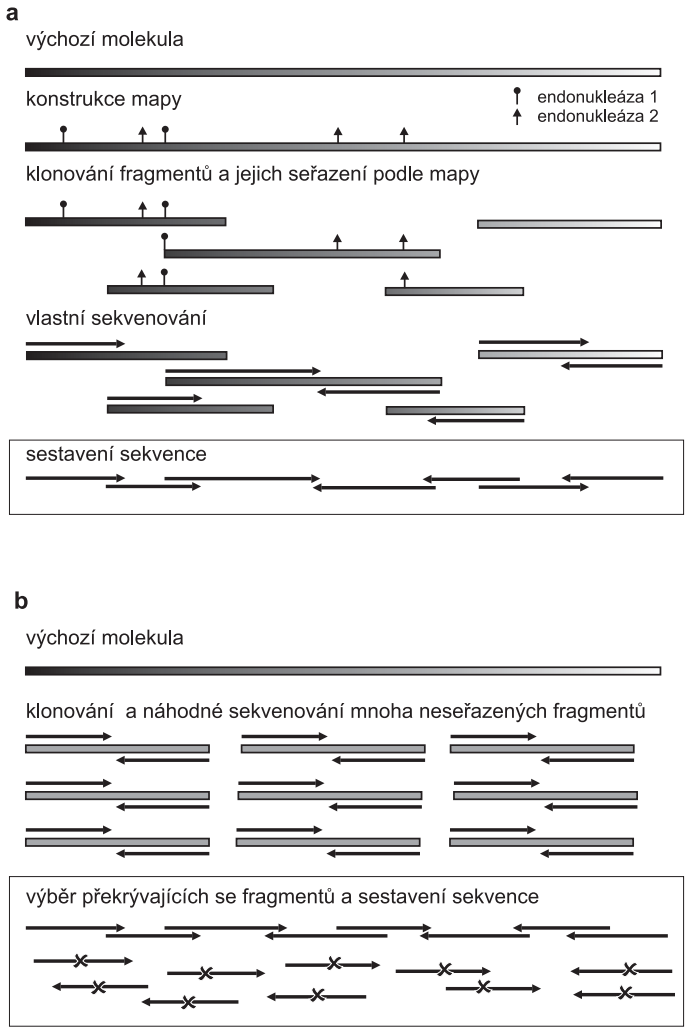
- Pro každou sekvenci obsahující HSP program stanoví úhrnnou normalizovanou hodnotu podobnosti ze všech HSP a tu pak použije pro statistické hodnocení. Výstup programu obsahuje seznam sekvencí s nejlepšími HSP spolu s hodnotami očekávatelnosti E (*expectance*), které udávají očekávaný počet sekvencí stejné nebo lepší podobnosti s dotazem ve stejně velké databázi složené z náhodných sekvencí. V případě proteinů lze za průkazné zpravidla považovat hodnoty E nižší než 0,001.

Hlavním zdrojem zrychlení proti přesným algoritmům je první krok, který by se dal přirovnat k odfiltrování málo informativních slov, jako jsou třeba spojky a ukazovací zájmena, v textových vyhledávačích, jako je třeba Google. Program BLAST je natolik rychlý a relativně nenáročný, že může být provozován i na serverech samotných databází jako volně dostupná vyhledávací služba.

Heuristické postupy se však uplatňují i v řadě jiných situací, než je vyhledávání vzájemně podobných sekvencí v databázi. Nenahraditelnou úlohu dnes hrají například také při **sestavování delší souvislé sekvence DNA** z přímých výstupů sekvenování. Jedna sekvenovací reakce totiž v typickém případě přečte „jen“ několik stovek nukleotidů, což je pouze zlomek délky typického genu, natož pak genomu. Až do poloviny 90. let minulého století proto jediný možný postup stanovení sekvence genomu vyžadoval pečlivé rozdělení genomu do důkladně zmapovaných fragmentů známé pozice; vlastní sekvenování představovalo vlastně jen zlomek celkového potřebného úsilí (obr. 3). Zásadní metodický přelom však přineslo sekvenování genomu bakterie *Haemophilus influenzae* v laboratoři Craiga Ventera, zakladatele soukromého Ústavu pro výzkum genomu (*The Institute for Genome Research*). Tito autoři totiž vymysleli způsob, jak obejít náročné a pracné mapování fragmentů na genom: stačí totiž osekvenovat dostatečné množství náhodně zvolených fragmentů DNA a jejich následné zmapování svěřit specializovanému programu, který pro každý fragment stanoví oblasti překryvu s jinými a překrývající se fragmenty pospojuje [12]. Právě vytvoření dostatečně rychlého a spolehlivého heuristického algoritmu pro sestavování genomové sekvence z krátkých fragmentů bylo zcela klíčovým krokem, který umožnil, aby se ze sekvenování genomů mohla stát dnes již téměř rutinní záležitost.

Poslední uvedený příklad také velice pěkně ilustruje obecný trend posledních desetiletí vývoje molekulární biologie, kdy postupně narůstá význam často velice rafinovaných postupů analýzy dat *in silico*, tedy ve virtuálním světě teoretických počítačových modelů[10]), někdy i na úkor tradičních postupů biologických (*in vivo*) a biochemických (*in vitro*). Staromilci však nemusí mít obavy — zánik experimentální biologie rozhodně nehrozí. Naopak: studium sekvencí DNA a proteinů otvírá zcela zásadním způsobem prostor pro formulaci otázek, na něž lze hledat odpovědi pouze v další experimentální práci.

[10]) Z jazykového hlediska by bylo správné používat obrat *in silicio*, protože křemík je silicium, nikoli silicum; ujalo se však *in silico* podle vzoru *in vivo* a *in vitro*. Sydney Brenner v jednom ze svých úvodníků v časopise Current Biology navrhuje termín „perverzní genetika“: v klasické genetice postupujeme od znaku ke genu, v genetice reverzní od genu ke znaku, a v té perverzní už za nás všechno dělá počítač.



Obr. 3. Klasická strategie sekvenování dlouhé molekuly DNA (a) a moderní postup využívající zásadním způsobem bioinformatické metody (b). Značky na mapě představují oblasti rozpoznávané restrikčními endonukleázami (viz legenda k obr. 2), vodorovné šipky pozici úseků odpovídajících digitální sekvenci získané např. metodou podle obr. 2. Kroky v rámečku se odehrávají ve virtuálním světě (*in silico*); přeškrtnuté šipky znamenají „nadbytečné“ sekvence, které nebyly využity při sestavování dlouhého úseku (i když mohou být v dalším kroku použity k ověření nebo zpřesnění sekvence). Schéma klasické metody je poněkud zjednodušené, protože ve skutečnosti by se mělo dbát na to, aby každý úsek byl sekvenován v obou směrech.

Poděkování. Bioinformaticky zaměřený výzkum v laboratoři autorky je podporován z prostředků Výzkumného centra MŠMT LC06004 „Integrovaný výzkum rostlinného genomu“.

L i t e r a t u r a

[1] Altschul, S. F., Gish, W. a kol.: *Basic local alignment search tool.* J. Mol. Biol. *215* (1990), 403–410.

[2] Altschul, S. F., Madden, T. L. a kol.: *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res. *25* (1997), 3389–3402.
[3] Baxevanis, A. D.: *Bioinformatics and the internet.* In: Baxevanis, A. D., Ouelette, B. F. F. (Ed.): *Bioinformatics. A practical guide to the analysis of genes and proteins.* Wiley-Interscience, New York, 2001, 1–17.
[4] Baxevanis, A. D., Ouelette, B. F. F. (ed.): *Bioinformatics. A practical guide to the analysis of genes and proteins.* Wiley-Interscience, New York 2001.
[5] Brenner, S.: *The world of genome projects.* BioEssays *18* (1996), 1039–1042.
[6] Chimpanzee Sequencing and Analysis Consortium: *Initial sequence of the chimpanzee genome and comparison with the human genome.* Nature *437* (2005), 69–87.
[7] Coen, E.: *The art of genes: How organisms make themselves.* Oxford University Press, Oxford 1999.
[8] Counsell, D.: *The Bioinformatics FAQ.* `http://bioinformatics.org/faq/`, 2004.
[9] Cvrčková, F.: *Úvod do praktické bioinformatiky.* Academia, Praha 2006.
[10] Cvrčková, F., Markoš, A.: *Beyond bioinformatics: can similarity be measured in the digital world?* J. Biosem. *1* (2005), 87–105.
[11] Felsenstein, J.: *Inferring phylogenies.* Sunderland, MA 2004.
[12] Fleischmann, R. D., Adams, M. D. a kol.: *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.* Science *269* (1995), 496–512.
[13] Gish, W., States, D. J.: *Identification of protein coding regions by database similarity search.* Nature Genetics *3* (1993), 266–272.
[14] Goff, S. A., Ricke, D. a kol.: *A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica).* Science *296* (2002), 92–100.
[15] Goffeau, A., Barrell, B. G. a kol.: *Life with 6000 genes.* Science *274* (1996), 563–567.
[16] Hogben, L.: *Matematika pro každého.* Borový, Praha 1948.
[17] Jacob, F.: *Evolution and tinkering.* Science *196* (1977), 1161–1166.
[18] Katrnoška, F., Křížek, M.: *Genetický kód a teorie monoidů.* PMFA *48* (2003), 207–222.
[19] Khorana, H. G.: *Nucleic acid synthesis in the study of the genetic code. Nobel Lecture, December 12, 1968., in: Nobel Lectures, Physiology or Medicine 1963–1970.* Elsevier, Amsterdam 1972.
[20] Markoš, A.: *Tajemství hladiny — hermeneutika živého.* Vesmír, Praha 2000; 2. vydání Dokořán, Praha 2003.
[21] Mouse Genome Sequencing Consortium: *Initial sequencing and comparative analysis of the mouse genome.* Nature *420* (2002), 520–562.
[22] Nirenberg, M.: *The genetic code. Nobel Lecture, December 12, 1968.* In: *Nobel Lectures, Physiology or Medicine 1963–1970*, Elsevier, Amsterdam 1972.
[23] Novák, D.: *Příspěvek elektronické konference Záhady, podrubrika Psychotronika.* `http://www.zahady.cz/`, 2001.
[24] Pearson, W. R.: *Protein sequence comparison and protein evolution: Tutorial — ISMB 2000.* `http://www.people.virginia.edu/~wrp/`, 2001.
[25] Pearson, W. R., Lipman, D. J.: *Improved tools for biological sequence comparison.* Proc. Natl. Acad. Sci. U. S. A. *85* (1988), 2444–2448.
[26] The Arabidopsis Genome Initiative: *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature *408* (2000), 796–815.
[27] The International Human Genome Sequencing Consortium: *Initial sequencing and analysis of the human genome.* Nature *409* (2001), 860–921.
[28] Venter, J. C., Adams, M. D. a kol.: *The sequence of the human genome.* Science *291* (2001), 1304–1351.
[29] Yu, J., Hu, S. a kol.: *A draft sequence of the rice genome (Oryza sativa L. ssp. indica).* Science *296* (2002), 79–92.
[30] Zrzavý, J., Storch, D., Mihulka, S.: *Jak se dělá evoluce.* Paseka, Praha 2004.