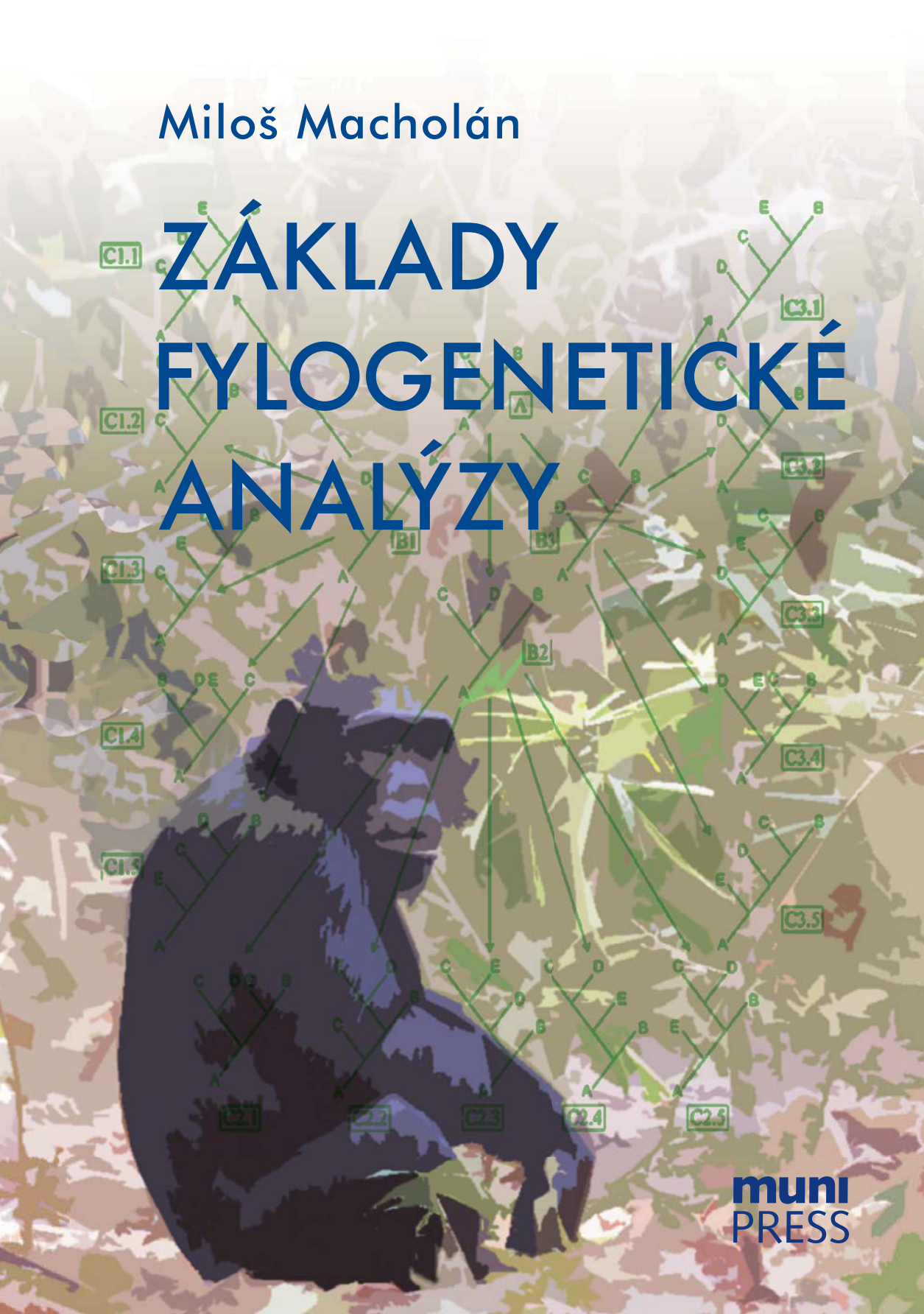
Miloš Macholán

# ZÁKLADY FYLOGENETICKÉ ANALÝZY




muni PRESS

# ZÁKLADY FYLOGENETICKÉ ANALÝZY

Miloš Macholán

muni
PRESS

# ZÁKLADY FYLOGENETICKÉ ANALÝZY

Miloš Macholán

MASARYKOVA UNIVERZITA
BRNO 2014

KATALOGIZACE V KNIZE – NÁRODNÍ KNIHOVNA ČR
Macholán, Miloš
Základy fylogenetické analýzy / Miloš Macholán. – Vyd. 1. –
Brno : Masarykova univerzita, 2014. – 289 s.
Anglické resumé
ISBN 978-80-210-6363-1

575.86 * 543.06
– fylogeneze
– analytické metody
– přehledy

575 - Obecná genetika. Obecná cytogenetika. Evoluce [2]

Citace
MACHOLÁN, Miloš. *Základy fylogenetické analýzy*. Brno: Masarykova univerzita, 2014, 289 s.
ISBN 978-80-210-6363-1.
DOI: 10.5817/CZ.MUNI.M210-6363-2014

*Knihu recenzoval*
*prof. RNDr. Karol Marhold, CSc.*


ISBN 978-80-210-7712-6 (online : pdf)
ISBN 987-80-210-6363-1 (brožovaná vazba)
DOI: 10.5817/CZ.MUNI.M210-6363-2014

# Obsah

# Předmluva

Pravděpodobně veškerý život na této planetě je spojen společným původem. Odhalení vzájemných fylogenetických vztahů mezi všemi organismy, tzv. stromu života, je jedním z předních úkolů evoluční biologie. V poslední době jsme svědky revolučních pokroků v molekulárněgenetických metodách, především v sekvenování nejen jednotlivých genů či jejich částí, ale i celých genomů. Na druhé straně mohutný rozvoj výpočetní techniky a tvorba sofistikovaného softwaru nám umožňuje zpracování obrovského – a stále rychleji rostoucího – množství dat. Tyto pokroky umožnily konstrukci přesnějších fylogenetických stromů a genových genealogií a ty zase zpětně umožňují odhalit a pochopit celou řadu biologických procesů na různých úrovních hierarchie života. Fylogenetické metody tak přestávají být doménou pouze systematicky zaměřených biologů, ale stávají se stále více nedílnou součástí genetiky, biogeografie, studia ontogeneze, chování, ekologie, epidemiologie, ochranářské biologie a studia evoluce vůbec.

Rozšiřující se možnosti při vyvozování fylogenetických vztahů však s sebou nesou i stále větší důraz na výběr správné metody, vyplývající ze znalosti podstaty používaných metod a většinou alespoň elementárního chápání procesů a mechanismů, které evoluci použitých znaků anebo fylogenezi analyzovaných organismů podmiňují. Bohužel, stále větší dostupnost kvalitního hardwaru a softwaru často svádí k postupu založenému na principu „vlož data – stiskni knoflík – publikuj výsledky“, aniž by bylo zřejmé, proč byla použita ta která metoda a jakým způsobem lze získané výsledky (případně rozdíly mezi nimi) interpretovat. Tato kniha přináší přehled a stručný popis v současnosti nejpoužívanějších metod fylogenetické analýzy. Měla by poskytnout základní informaci o tom, z jakých předpokladů jednotlivé metody vycházejí, v čem spočívají jejich výhody a nevýhody a v čem tkví rozdíly mezi nimi. Pro většinu těchto metod jsou podány základní matematické vztahy, které však může nematematicky zaměřený čtenář přeskočit, aniž by ztrácel orientaci v textu. Přesto doporučuji alespoň některé z nich projít – to se týká především notoricky známých metod jako například výpočet genetických distancí a metod shlukové analýzy. Mým cílem je poskytnout dostatečnou elementární informaci tak, aby čtenář byl schopen činit poučená rozhodnutí při výběru optimálního přístupu při řešení daného problému a pro daný typ dat, stejně jako získané výsledky správně interpretovat. Přestože ke konstrukci fylogenetických stromů můžeme použít širokou škálu znaků, zde se zaměřím hlavně na znaky molekulární, především na sekvence DNA. Ostatní typy dat budou zmíněny jen okrajově, a to pouze tam, kde to bude nutné, s výjimkou morfometrických znaků, kterým je věnována samostatná kapitola. Závěrečná část pojednává o koalescenci, která nachází uplatnění především na populační úrovni, a proto se z rámce fylogenetických metod poněkud vymyká, nicméně jsem přesvědčen, že znalost alespoň základních principů této teorie je velice důležitá.

Tento přehled samozřejmě není a ani nemůže být vyčerpávající. Snažil jsem se zaměřit pouze na nejznámější metody – podrobnější informace je třeba vyhledat ve specializované literatuře. Také použité literární odkazy jsem se snažil omezit na minimum. Kromě odborných článků a kapitol v knihách jsem čerpal z následujících monografií: Wiley (1981),

Wiley et al. (1991), Harvey a Pagel (1993), Hillis et al. (1996), Marcus et al. (1996), Hartl a Clark (1997), Kitching et al. (1998), Page a Holmes (1998), Nei a Kumar (2000), Wiens (2000), Gillespie (2001), Hall (2001), MacLeod a Forey (2002), Marhold a Suda (2002), Salemi a Vandamme (2003), Felsenstein (2004), Zelditch et al. (2004), Balding et al. (2007). Nejobsáhlejším a nejpodrobnějším přehledem z nich je zřejmě monografie J. Felsensteina *Inferring Phylogenies* z roku 2004. Fylogenetickými metodami se zabývají i další knihy: Kitching et al. (1993), Scotland et al. (1994), Martins (1996), Pääbo (1998), Page (2002), Albert (2006), Gascuel a Steel (2003), Semple a Steel (2003), Gascuel (2007), Xia (2007), Lemey et al. (2009), Knowles a Kubatko (2010), Kuo et al. (2014).

Na fylogenetiku můžeme pohlížet ze dvou odlišných úhlů. Na jedné straně především zastánci fylogenetické taxonomické školy (tzv. kladistiky) vidí rekonstrukci fylogeneze jako *filozofický* problém. Podle tohoto názoru vzhledem k tomu, že skutečnou historii studovaných organismů nemůžeme až na vzácné výjimky nikdy poznat, nemá smysl nějakou „pravdu" o fylogenezi hledat a jediné, co můžeme udělat, je snažit se pozorovaná data nějakým srozumitelným způsobem sumarizovat. Při tom se musíme spolehnout na nějaké (pokud možno co nejjednodušší) kritérium, pevný bod, na kterém můžeme dále stavět. Tímto kritériem je princip Ockhamovy břitvy, podle kterého je preferována jednodušší hypotéza, předpokládající menší počet evolučních kroků, před složitějšími, předpokládajícími větší počet změn. Navíc každá historie obsahuje určitou míru nahodilosti, která už z principu neumožňuje v ní hledat nějaké zákonitosti, jež by bylo možno popisovat pomocí jednoduchých modelů. Naproti tomu jiní vidí rekonstrukci fylogeneze jako *statistický* problém, nikoli jako filozofický světonázor, skrze který posuzujeme svět okolo sebe. Aniž bych měl v úmyslu zasahovat do polemik mezi oběma tábory, v tomto textu budu vycházet z druhého úhlu pohledu, protože bez ohledu na svoje osobní preference se domnívám, že pouze tak lze čtenáři předložit široké spektrum v současnosti nejčastěji používaných fylogenetických metod.

Na tomto místě je třeba poděkovat především prof. RNDr. Karolu Marholdovi, CSc., Mgr. Natálii Martínkové, Ph.D., a Mgr. Peteru Mikulíčkovi, Ph.D., za cenné připomínky k rukopisu knihy. V neposlední řadě děkuji Nakladatelství Masarykovy univerzity v čele s PhDr. Alenou Mizerovou, že mi umožnilo tuto knihu vydat.

Brno, červenec 2014 Miloš Macholán

# 1. Úvod

Základem fylogenetické analýzy je datový soubor, který má formu matice stavů znaků (matice dat), zkráceně označované jako data. Řádky matice obsahují informaci o jednotlivých objektech, které budeme obecně nazývat **taxony** (někdy označované jako operační taxonomické jednotky, OTU). Pod tímto pojmem si můžeme představit populace, virové kmeny, druhy, popř. vyšší taxonomické jednotky (rody, řády, třídy, kmeny atd.). Každý sloupec matice představuje jeden znak (*character*). V případě molekulárních dat je vstupní matice tvořena souborem příslušným způsobem seřazených sekvencí DNA (případně sekvencí aminokyselin), kde znaky představují jednotlivé pozice (*sites*) v sekvenci a konkrétní báze (aminokyselina) představují stavy znaku (*character states*). Například jestliže na 362. pozici sekvence genu pro cytochrom *b* je u šimpanze přítomen adenin (A), u bonoba guanin (G) a u člověka thymin (T), je tato pozice znakem a A/G/T jsou stavy tohoto znaku pro příslušné taxony.[1] Kromě nukleotidů (resp. bází) nebo aminokyselin mohou elementy matice představovat např. binární kódy, přítomnost/absence znaku, naměřené kvantitativní hodnoty atd.

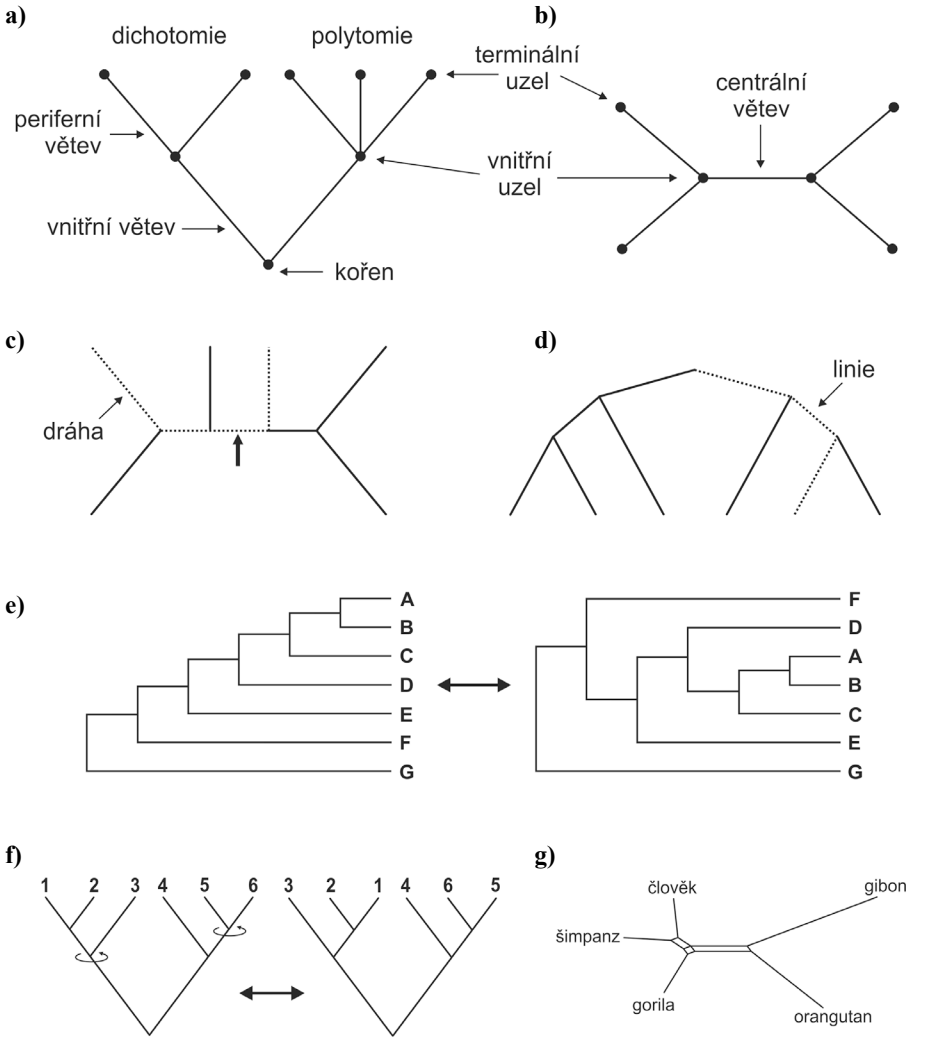
## DEFINICE ZÁKLADNÍCH POJMŮ

Grafickým vyjádřením fylogeneze je **fylogenetický strom** neboli **fylogenie** (*phylogeny*). Za strom můžeme považovat jakýkoli necyklický graf spojující bez přerušení jednotlivé **uzly** (*nodes*, *vertices*). Uzel může být buď **terminální** neboli **externí** (*terminal node*, *external node*, *leaf*, *pendant vertex*), nebo **vnitřní** (*internal node*, *internal vertex*). Stromy, jejichž terminální uzly jsou pojmenovány, se nazývají **označené** (*labeled*). Vnitřní uzly někdy nazýváme hypotetické taxonomické jednotky, HTU.

Jednotlivé uzly jsou spojeny **větvemi** (*branches*, *edges*). Větve končící terminálním uzlem jsou **periferní** (*peripheral branches*, *peripheral edges*), větve spojující dva vnitřní uzly jsou **vnitřní** (*internal branches*, *internal edges*) a konečně větev spojující čtyři periferní větve se označuje jako **centrální** (obr. 1.1a, b). Jako **sousední** (přilehlé) se označují větve, které sdílejí společný uzel. Sekvence sousedních větví $v_1$, $v_2$, $v_3$, …, $v_n$ se nazývá **dráha** (*path*). Každý fylogenetický strom je charakterizován určitým prostorovým uspořádáním jednotlivých větví čili **topologií**. Jednotlivým větvím jsou často přiřazeny určité hodnoty – váhy (*weights*) neboli **délky** (*lengths*). Ve fylogenetickém kontextu těmito vahami může být čas, počet změn (např. substitucí) nebo jejich pravděpodobnost.

Strom může být buď **s kořenem** (*rooted tree*), nebo **bez kořene** (*unrooted tree*). Kořen stromu představuje společného předka všech ostatních taxonů (tj. vnitřních i vnějších uzlů stromu). Ve druhém případě není tento předek (uzel) identifikován. Strom s kořenem dostaneme přiřazením kořene buď jednomu z uzlů (vnitřnímu nebo terminálnímu),

[1] Tato teminologie není univerzální, např. v některých kladistických publikacích se místo pojmu „znak“ používá termín „transformační série“ a termín „znak“ je pak chápán ve významu konkrétního „stavu“.




**Obr. 1.1** Příklad jednoduchého stromu s kořenem (a) a bez kořene (b) a vysvětlení některých základních pojmů. Fylogenetická dráha (c) je součtem všech větví mezi dvěma terminálními uzly, kdežto linie (d) spojuje jeden terminální uzel s kořenem stromu; zatímco dráhy můžeme identifikovat na stromech s kořenem i bez kořene, linie existují pouze na stromech s kořenem. Jednotlivé části stromu mohou být rotovány, aniž by se změnil jeho smysl, a proto dvojice stromů na obr. e) a f) jsou totožné. Na obr. g) je ukázka fylogenetické sítě pěti druhů hominoidů.

nebo vložením uzlu nového (např. ze stromu 1.1c dostaneme strom 1.1d vložením uzlu v místě označeném šipkou). Jestliže se ve vnitřním uzlu stýkají pouze tři větve, hovoříme o **bifurkaci** neboli **dichotomii**, v případě více větví jde o **multifurkaci** neboli **polytomii** (obr. 1.1a). Strom, ve kterém se vyskytují pouze bifurkace, se označuje jako plně vyřešený čili **binární** (*fully resolved*, *binary*) nebo též **striktně bifurkační**. Strom, obsahující jediný vnitřní bod, se nazývá **hvězdicový** (*star tree*) (obr. 1.2). Je to strom, kde minimálně všechny terminální uzly jsou pojmenovány (*labeled*).

Kořen fylogenetického stromu můžeme stanovit přidáním jedné nebo více **vnějších skupin** (*outgroups*). Volba vnější skupiny nebo skupin může být klíčová pro správné určení kořene. Tento krok není triviální, protože vnější skupina by neměla být od studovaných taxonů (tvořících tzv. **vnitřní skupinu**, *ingroup*) fylogeneticky příliš vzdálená (nemá např. smysl hledat kořen stromu blízce příbuzných druhů primátů přidáním sekvence žraloka), ale ani příliš blízká, jinak bychom riskovali, že je ve skutečnosti součástí vnitřní skupiny, místo aby stála mimo ni. V ideálním případě by tedy měla tvořit vůči zkoumané skupině tzv. skupinu sesterskou (např. na obr. 1.1e je taxon G sesterský skupině taxonů A–F).

Terminální uzly mohou směřovat kterýmkoli směrem (srv. stromy 1.1a, 1.1d a 1.1e), zpravidla jsou však orientovány doprava nebo nahoru. Na orientaci stromu tedy nezáleží. Dokonce i jednotlivé části stromu mohou být rotovány, aniž by se změnil jeho smysl. Například stromy na obr. 1.1e–f jsou totožné. Strom bez kořene býval dříve nesprávně označován jako **síť** (*network*), avšak tento termín má podle matematické definice jiný význam, neboť v síti dochází nejen ke štěpení větví, ale i k jejich spojování (obr. 1.1i). V současné době se s těmito sítěmi setkáváme zejména ve fylogeografických studiích, kde jsou tímto způsobem znázorněny alternativní, co do počtu kroků (např. substitucí) ekvivalentní spojení jednotlivých objektů (zpravidla haplotypů). Sítí však můžeme znázornit i případy skutečného splývání dvou linií, například studujeme-li evoluci lidských jazyků.

## Kolik existuje možných fylogenetických stromů?

Nejmenší počet taxonů tvořících fylogenetický strom jsou tři. Jestliže tento strom nemá kořen, existuje pouze jediná varianta, pokud kořen má, jsou možné varianty tři. Přidáme-li další taxon, můžeme vytvořit tři alternativní stromy bez kořene – pokud identifikujeme kořen, pro každý ze tří původních alternativ můžeme vytvořit pět různých stromů, celkem tedy 15. S rostoucím počtem taxonů ovšem množství možných výsledků dramaticky roste. Máme-li $n$ taxonů, je počet možných stromů bez kořene dán vztahem (Felsenstein 1978b):

$$\frac{(2n-5)!}{2^{n-3}(n-3)!} \tag{1.1}$$

a počet stromů s kořenem bude

$$\frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{1.2}$$

Výsledky pro různé počty taxonů jsou uvedeny v tabulce 1.1.

| Taxony | Stromů bez kořene | Stromů s kořenem |
|---|---|---|
| 3 | 1 | 3 |
| 4 | 3 | 15 |
| 5 | 15 | 105 |
| 6 | 105 | 945 |
| 7 | 945 | 10 395 |
| 8 | 10 395 | 135 135 |
| 9 | 135 135 | 2 027 025 |
| 10 | 2 027 025 | 34 459 425 |
| 11 | 34 459 425 | 654 729 075 |
| 12 | 654 729 075 | 13 749 310 575 |
| 13 | 13 749 310 575 | 316 234 143 225 |
| 14 | 316 234 143 225 | 7 905 853 580 625 |
| 15 | 7 905 853 580 625 | 213 458 046 676 875 |
| 20 | 213 458 046 676 875 | 8 200 794 532 637 891 559 375 |
| 30 | 8 200 794 532 637 891 559 375 | $4{,}9518\times10^{38}$ |
| 40 | $4{,}9518\times10^{38}$ | $1{,}00986\times10^{57}$ |
| 50 | $1{,}00986\times10^{57}$ | $2{,}75292\times10^{76}$ |

**Tabulka 1.1** Počet stromů s kořenem a bez kořene pro různý počet taxonů.

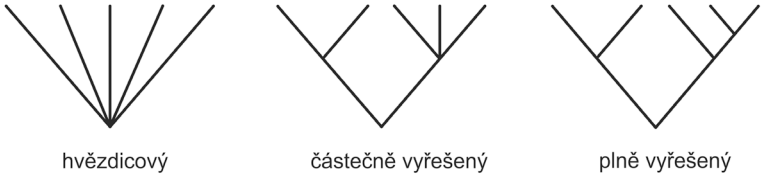
Pro více než 10 taxonů je velmi obtížné prozkoumat všechny stromy s kořenem. Při počtu taxonů přesahujícím 30 množství stromů s kořenem mnohonásobně převyšuje hodnotu Avogadrovy konstanty[2] a prozkoumání všech už není možné. Pro 50 taxonů výsledný počet všech možných stromů s kořenem dokonce přesahuje počet elektronů ve viditelném vesmíru (tzv. Eddingtonovo číslo).

## TYPY DAT

Všechna data používaná při fylogenetické analýze spadají do jedné ze dvou základních kategorií. První kategorií jsou jednotlivé znaky, druhou data ve formě vzdáleností (distancí) či podobností. „Jednotlivými" rozumíme jak znaky kvalitativní (diskrétní), tak kvantitativní. Zatímco data ve formě znaků mohou být převedena na distance/podobnosti, opačný postup není možný. Je třeba si uvědomit, že některé metody, například měření imunologické reakce na cizí antigeny nebo DNA-DNA hybridizace poskytují výsledky výhradně ve formě distancí mezi dvojicemi taxonů.

Abychom mohli data využít pro konstrukci fylogenetického stromu, je zpravidla nutno splnit některé podmínky. Jednou z nich je podmínka **nezávislosti** jednotlivých znaků. Pokud mezi jednotlivými znaky existuje vzájemná závislost, musíme brát v úvahu kovariance mezi nimi, což následné analýzy značně komplikuje. Další podmínkou je **homologie** daného znaku mezi srovnávanými taxony. V případě molekulárních znaků je nutno zkoumat sekvence vzájemně **ortologní**, tj. sekvence zděděné dvěma druhy od společného

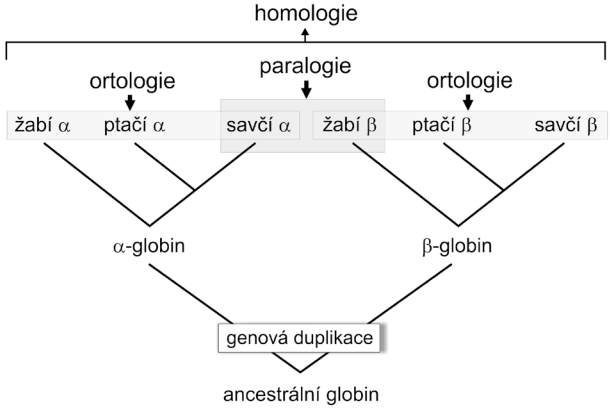
[2] Avogadrova konstanta vyjadřuje počet částic v jednotkovém látkovém množství; její hodnota je $6{,}022 \times 10^{23}$ mol$^{-1}$.




**Obr. 1.2** Ukázka tří stromů s různou mírou objasnění jednotlivých větvení.

předka (např. sekvence genu pro podjednotku alfa hemoglobinu u druhů a a B). Ortologní sekvence mohou být někdy nechtěně zaměněny za sekvence **paralogní**. Tak se označují sekvence genů, které vznikly genovou duplikací (Ohno 1970), například sekvence genů pro podjednotku alfa a beta hemoglobinu (obr. 1.3). Jinými slovy, zatímco ortologní sekvence divergují po speciační události (tj. po vzniku nových druhů), paralogní sekvence divergují po události duplikační. Paralogní geny vzniklé duplikací celého genomu se někdy označují jako ohnologní (Wolfe 2000.) Podobné, byť ne tak časté riziko představují sekvence **xenologní**, které byly do zkoumaného genomu vneseny z jiného organismu. Určitou variací paralogních a xenologních sekvencí jsou pseudogeny přenesené do jaderného genomu z mitochondriální DNA (anglicky se zpravidla označují zkratkou **numt** z *nuclear mitochondrial sequence*). Tyto pseudogeny mohou být i velmi velké, například u některých kočkovitých šelem dosahuje jejich velikost přibližně 12,5 kb (Kim et al. 2006) a obsahuje tudíž převážnou část mitochondriálního genomu. S rostoucím množstvím molekulárních dat se ukazuje, že „numity" mohou být značně rozšířené (viz www.pseudogene.net) a nebezpečí jejich záměny s mitochondriálními sekvencemi je tedy potenciálně velké.

Znaky mohou být kvantitativní nebo kvalitativní. Kvalitativní znaky mohou být buď **binární**, mající pouze dva možné stavy, nebo **vícestavové** (*multistate characters*) se třemi a více možnými stavy. Vícestavové znaky mohou být **seřazené** (*ordered*), nebo **neseřazené** (*unordered*). Typickým příkladem neseřazených vícestavových znaků jsou sekvence




**Obr. 1.3** Vztah ortologních a paralogních sekvencí na příkladu α- a β-globinových genů.

nukleotidů, protože zpravidla není žádný apriorní důvod předpokládat, že daný nukleotid vznikl ze zcela konkrétního nukleotidu nebo že určitá nukleotidová báze je intermediární mezi dvěma jinými bázemi. Pojem *pořadí* znaků, definující možnou (povolenou) transformaci jednoho stavu znaku v druhý (např. ze stavu *A* může vzniknout stav *B* nebo naopak, nikoli však stav *C*), bychom však neměli zaměňovat za *polaritu* znaků, která popisuje směr evoluce stavů znaku (např. v evoluci taxonu X došlo ke změně znaku v sekvenci stavů *A*-*B*-*C*, nikoli *C*-*B*-*A*).

## Nukleotidové a proteinové sekvence

S rozvojem a zefektivněním molekulárních technik jsou sekvence nukleotidů nebo aminokyselin stále častějším zdrojem dat pro fylogenetickou analýzu. Především sekvence DNA dnes díky novým technologiím ve fylogenetice naprosto dominují. Tyto metody, souhrnně označované jako sekvenování příští generace, častěji zkratkou NGS (*next generation sequencing*), jsou schopny za jeden den získat více než tisíckrát větší počet párů bází než nejvýkonnější klasické sekvenátory. V současnosti existuje několik systémů, které se metodicky poněkud liší – mezi nejznámější patří „pyrosekvenování" metodou 454 (Roche), Solexa (Illumina) nebo SOLiD (Life Technologies) –, všechny jsou však založeny na simultánním sekvenování obrovského množství velmi krátkých fragmentů DNA. Jednotlivé osekvenované fragmenty jsou nakonec sestaveny do výsledné sekvence. Sestavení celé sekvence z velkého množství malých fragmentů pochopitelně vyžaduje výkonný počítač. Technický vývoj stále pokračuje a už dnes existují nebo jsou vyvíjeny metody, které poskytují vyšší výtěžek než NGS a současně snižují náklady na sekvenování (*3rd generation sequencing*). Stále častěji jsou proto fylogenetické analýzy založeny na sekvencích mnoha genů. Taková data se označují jako fylogenomická a pro oblast fylogenetiky, která je na nich založená, se používá termín **fylogenomika**.

Na první pohled je využití sekvencí koncepčně poměrně jednoduché. Podmínka homologie však v tomto případě znamená nejen to, že zkoumáme ortologní sekvence, ale také že nukleotidy na dané pozici u všech zkoumaných taxonů odvozují svůj původ od této pozice u jejich společného předka. Vzhledem k výskytu delecí a inzercí během evoluce je stanovení této homologie někdy poměrně obtížné (zejména při porovnávání vzdáleně příbuzných taxonů).

## Restrikční data

Restrikční endonukleázy (restriktázy) jsou enzymy štěpící DNA v místě specifické krátké sekvence, tzv. restrikčního místa (*restriction site*), nebo v jeho bezprostřední blízkosti. Pozice těchto míst na restrikční mapě lze považovat za znaky a přítomnost nebo absenci těchto míst u daného taxonu za stavy znaku (*restriction-site data*). V případě, že restrikční mapu nemáme k dispozici, lze za stav znaku považovat přítomnost/absenci restrikčního fragmentu o příslušné délce (*restriction-fragment data*). Přestože donedávna byl tento druhý typ restrikčních dat při fylogenetických analýzách často používán, především jako tzv. RFLP (*restriction fragment length polymorphisms*, polymorfismus délky restrikčních fragmentů), nelze je příliš doporučit. Důvodem je skutečnost, že předpoklad nezávislosti jednotlivých znaků v tomto případě není splněn. Jestliže uvnitř fragmentu mezi dvěma

restrikčními místy vznikne místo třetí, původní delší fragment se rozdělí na dva kratší. Tím pádem nebudou mít dva druhy, sdílející dvě ze tří restrikčních míst, žádný společný fragment, což může být zdrojem dosti vážných potíží. Obdobným problémem je možnost změny délky fragmentu inzercí nebo delecí úseku DNA, které povedou k chybnému závěru, že daný taxon příslušný fragment postrádá, přestože má obě homologní restrikční místa.

Posledním problémem, charakteristickým pro oba typy restrikčních dat, je asymetrie pravděpodobnosti vzniku a zániku restrikčního místa. Představme si sekvenci složenou ze 6 bp, ze které se substitucí na jedné z pozic stane restrikční místo. V tomto případě existuje pouze jedna z 18 možných substitucí (šest míst, tři možné typy záměn jednoho nukleotidu za druhý), kterou může nové restrikční místo vzniknout. Naopak jestliže daná šestinukleotidová sekvence už restrikčním místem je, potom záměna *kteréhokoli* nukleotidu znamená, že restrikční enzym tuto sekvenci nerozpozná. Ztráta restrikčního místa je proto mnohem pravděpodobnější než jeho vznik. Tento argument neznamená, že nelze porovnávat ztráty restrikčních míst mezi taxony během evoluce, pouze ukazuje na nutnost speciálního zacházení s restrikčními daty. V současnosti se už RFLP data pro fylogenetickou analýzu prakticky nepoužívají. Štěpení pomocí DNA restriktáz však využívají některé pokročilejší metody, například RADSeq nebo AFLP (viz níže). **RADSeq** (*restriction site-associated DNA sequencing*; Baird et al. 2008; Davey a Blaxter 2010) patří do skupiny metod souhrnně nazývaných *reduced-representation genome sequencing* a je kombinací restrikčního štěpení a NGS. Genomová DNA jedince je nejprve štěpena pomocí restriktázy na menší fragmenty, na které je napojen adaptor P1 obsahující primer pro amplifikaci, primer pro NGS a tzv. molekulární identifikátor (*molecular identifier*, MID), který umožňuje bezpečnou identifikaci daného jedince. Fragmenty DNA různých jedinců jsou posléze zkombinovány do jednoho vzorku a náhodně zkráceny na menší bloky, dlouhé několik set párů bází. Na ty je napojen druhý adaptor (P2), speciálně uzpůsobený tak, že se na něj nenapojí zpětný primer, dokud nedojde k prvnímu kolu amplifikace od pozice P1 primeru. Tím je zajištěno, že všechny fragmenty vytvořené konečnou PCR reakcí obsahují P1 adaptor, MID, částečné restrikční místo, několik set párů bází (200–500) přiléhajících k tomuto místu a P2 adaptor. Tyto fragmenty jsou sekvenovány pomocí některé z metod NGS (první protokoly byly vytvořeny pro systém Illumina). Nakonec jsou sekvence každého jedince odděleny na základě jejich MID. Metoda RADSeq může být použita i k detekci polymorfismu restrikčních míst (přítomnost/absence), jednonukleotidových rozdílů (*single-nucleotide polymorphisms*, SNP) nebo delecí/inzercí. Sekvenování pouze části genomu, vymezené přítomností restrikčních míst, umožňuje rychlé a relativně levné získání velkého množství dat pro fylogenetickou analýzu (Rubin et al. 2012).

## Alozymy

Dalším typem dat, jejichž role ve fylogenetice je dnes již víceméně historická, jsou alozymy. Alozymová data jsou většinou prezentována ve formě relativních četností (frekvencí) jednotlivých alel na každém z analyzovaných lokusů ve zkoumaných populacích nebo taxonech. Původní a dosud převládající způsob zpracování tohoto typu dat je založen na výpočtu matice genetických distancí. S rozvojem metod založených na přímé analýze znaků se objevily snahy alozymová data kódovat do diskrétní podoby. V zásadě existují tři způsoby kódování, přičemž každý z nich je problematický.

Prvním a historicky nejstarším způsobem kódování alozymových dat je považovat každou alelu za jeden znak a buď její přítomnost/absenci, nebo její frekvenci za stav tohoto znaku. Tento postup však trpí stejným problémem jako restrikční fragmenty, neboť není splněn předpoklad nezávislosti znaků: protože součet frekvencí alel musí být vždy roven jedné, jakmile roste frekvence jedné alely, frekvence ostatních alel na tomtéž lokusu musí zákonitě klesat. Jestliže takto kódovaná data dále zkoumáme metodou maximální úspornosti (viz kap. 2), často dospějeme k závěru, že společný předek neměl žádné alely, nebo součet četností těchto alel se nerovná jedné.

Vzhledem k těmto problémům byla navržena další metoda kódování, kde znakem je lokus a stavem znaku je konkrétní alela (např. jeden taxon má jen alelu *a* a druhý jen alelu *b*, zatímco v případě polymorfismu bychom kódovali třetí stav *ab*). Přes jistý pokrok má i tato metoda určité nedostatky. Za prvé, pokud se mezi taxony vyskytuje velké množství odlišných alel v různých kombinacích, může se počet jedinečných kombinací blížit počtu taxonů. Takovéto znaky (lokusy) potom budou zdrojem nepatrné nebo vůbec žádné informace. Východiskem by bylo stanovit posloupnost vzniku jednotlivých alel (seřazené znaky, viz str. 15), avšak tento postup bývá často subjektivní a arbitrární. Druhým problémem tohoto přístupu je silná náchylnost k chybě výběru (*sampling error*). Jestliže jsou na lokusu přítomny dvě alely a jedna z nich je velmi vzácná, díky omezenosti našeho vzorku s velkou pravděpodobností tuto alelu nezjistíme a lokus budeme chybně kódovat jako monomorfní. Dokonce i když bychom vyšetřili celou populaci a alelové četnosti vypočetli naprosto přesně, stěží můžeme považovat za rovnocenné dvě populace (1, 2), lišící se četnostmi svých alel $a_1 = 0{,}01$, $b_1 = 0{,}99$ a $a_2 = 0{,}99$, $b_2 = 0{,}01$.

Třetím způsobem, který se snaží brát v úvahu informaci obsaženou v alelových četnostech, je tzv. kvantitativní kódování, ve kterém znakem je opět lokus, stavem znaku je pak relativní četnost (frekvence) alel. Alelové četnosti ve dvou či více vzorcích jsou ve formě kontigenčních tabulek podrobeny příslušnému statistickému testu, který má rozhodnout, zda tyto vzorky pocházejí z jediné homogenní populace. Pokud však nemáme k dispozici velké vzorky, je síla těchto testů poměrně malá a neprůkaznost rozdílů ještě neznamená, že tyto vzorky jsou homogenní. Tyto problémy způsobují, že metody vyžadující kódování alozymových dat do diskrétní podoby by měly být používány pouze v případech nízké úrovně polymorfismu.

Problémům s kódováním se snaží vyhnout metody, které využívají alelové četnosti přímo, aniž by je bylo nutno převádět na diskrétní znaky. Výše zmíněné námitky totiž většinou vyplývají ze skutečnosti, že tento typ dat je ve své podstatě kvantitativní povahy a jeho převod do diskrétní podoby (kromě toho, že tímto postupem dochází ke ztrátě informace) postrádá biologické opodstatnění. Mezi metody založené na přímém zpracování alelových četností patří například techniky minimalizující celkové množství změn ve frekvencích alel a modifikace metody maximální věrohodnosti na kvantitativní data, vycházející z modelu Brownova pohybu (viz kap. 9).

## SINE a LINE

Bouřlivý rozvoj sekvenovacích technik sice umožňuje shromáždění obrovských datových souborů obsahujících sekvence stovek až desetitisíců genů, avšak naše současné výpočetní a metodické možnosti jejich analýzy za tímto vývojem stále zaostávají. Nejrozsáhlejší

fylogenomické soubory je proto nutno analyzovat pomocí suboptimálních modelů, které snižují, nebo dokonce anulují výhody, které velikost těchto datových souborů poskytuje. Z toho důvodu byly hledány alternativní zdroje fylogenetických znaků, založené na mutacích většího rozsahu. Takovými znaky jsou například SINE a LINE elementy.

Krátké a dlouhé vmezeřené repetitivní elementy, zkráceně SINE (*short interspersed elements*) a LINE (*long interspersed elements*), patří do velké skupiny retroelementů, tedy transpozabilních elementů, které se šíří pomocí přepisu do RNA a opětovného začlenění do řetězce DNA pomocí enzymu reverzní transkriptázy (SINE ve skutečnosti nejsou pravými retroelementy, protože nekódují vlastní reverzní transkriptázu a pro reverzní transkripci používají enzym kódovaný LINE). Nejznámějšími SINE jsou *Alu* sekvence u člověka a B1 a B2 elementy u myši. Replikované elementy jsou vřazeny náhodně v různých částech genomu, a jakmile se začlení do řetězce DNA, prakticky již z něho nemohou být odstraněny (výjimkou by mohla být jen delece velkého úseku DNA). Oba typy repeticí samozřejmě podléhají substitucím a inzercím/delecím, čímž dochází k jejich postupné divergenci, avšak v případě taxonů, jejichž doba divergence nepřesahuje 50 milionů let, je jejich použití při konstrukci fylogenetického stromu velmi výhodné. Na rozdíl od SINE, jejichž délka zpravidla nepřesahuje 400 bp, jsou LINE mnohem delší (několik tisíc bp). Délka jednotlivých kopií však velmi kolísá, a proto je práce s nimi obtížnější. Na druhou stranu díky své délce mohou poskytnout spolehlivý odhad doby divergence sekvencí. SINE a LINE elementy mohou potenciálně být zdrojem velkého množství specifických znaků, například LINE elementy L1 tvoří 17 % lidského genomu, SINE elementy *Alu* 12 %. Protože k inzerci retroelementů dochází v náhodných místech genomu a pravděpodobnost, že stejný element bude nezávisle začleněn přesně do stejného místa ve dvou a více liniích, je velmi nízká, jsou tyto znaky většinou považovány za prosty homoplazií (tj. analogických stavů vzniklých v důsledku konvergence, paralelismu nebo reverze k původnímu stavu). Nicméně jak ukázali Han et al. (2011), tento předpoklad může být, alespoň v některých případech, příliš optimistický.

## Pořadí genů

Fylogenetická analýza založená na strukturním uspořádání genů si v poslední době získává rostoucí oblibu. Někteří autoři dokonce tento typ dat považují za potenciálně více informativní vzhledem k tomu, že změny v pořadí genů jsou málo frekventované, a tudíž méně náchylné k homoplaziím, a protože je málo pravděpodobné, že by identické sekvence genů vznikly v různých liniích nezávisle. A tak přestože znaky ve formě pořadí genů většinou neumožňují konstrukci plně vyřešeného stromu, jejich užitečnost tkví především ve studiu velmi starých divergencí. Problém je v tom, že tyto znaky se nemohou vyvíjet nezávisle na sobě, protože jsou definovány právě prostřednictvím vztahů mezi sebou. Teprve budoucnost ukáže, zda budou vyvinuty metody, které by tento problém braly v úvahu.

## miRNA

Poměrně novou a dosud málo prozkoumanou skupinou znaků vhodných pro fylogenetickou analýzu je zvláštní třída RNA známá jako miRNA (*microRNA*). Zralé molekuly miRNA jsou velmi malé, přibližně 22 bp dlouhé jednovláknové řetězce nekódující RNA,

které se podílí na regulaci genové exprese prakticky u všech mnohobuněčných organismů. K jejich objevu došlo až v roce 1993, jejich název však pochází až z roku 2001. Jsou kódovány buď intronovými, nebo mezigenovými sekvencemi. Ty jsou nejprve přepsány do 70 bp dlouhého řetězce primární miRNA (pri-miRNA), která má podobu běžné mRNA s čepičkou na 5′-konci a poly-A na 3′-konci. Tento prekurzor se složí do tzv. vlásenkové (*hairpin*) struktury a je dále upraven odštěpením kolem devíti nukleotidů z jeho báze za vzniku prekurzorové miRNA (pre-miRNA). Ta cestuje z jádra do cytoplazmy, kde z ní činností endonukleázy *dicer* vzniká dvouvláknová a nakonec jednovláknová zralá miRNA (většinou je pro další funkci používán jen jeden z obou řetězců). Zralé, plně funkční miRNA jsou komplementární k části jedné nebo více mRNA – u živočichů zpravidla k oblasti 3′-UTR (*untranslated region*), u rostlin ke kódujícím oblastem. Navázání na tuto oblast způsobí buď zablokování translace příslušné mRNA, nebo její rozštěpení. Účinkem miRNA je tedy negativní regulace genů.

Proč jsou miRNA z hlediska fylogenetické analýzy zajímavé? Důvodů je několik: 1) nové miRNA můžeme v genomu identifikovat, aniž bychom předem znali jejich sekvenci; 2) časem v genomu vznikají stále nové rodiny těchto molekul; 3) jejich sekundární ztráta je vzácná; 4) frekvence substitucí je u nich nízká; 5) stejně tak je nízká i pravděpodobnost jejich konvergentní evoluce (Tarver et al. 2013). K typickým vlastnostem miRNA patří vysoká konzervativnost jejich primární sekvence. Protože je jejich evoluční původ vzájemně nezávislý, jsou brány jako znaky diskrétní, se stavy „přítomnost/absence“, které lze snadno kódovat a analyzovat pomocí standardních fylogenetických metod.

## Jiné typy molekulárních znaků

Spektrum typů dat použitelných pro fylogenetickou analýzu je velmi široké, ne všechny jsou však stejně vhodné. Dnes už prakticky nepoužívanou metodou získání molekulárních dat je **RAPD** (*randomly amplified polymorphic DNA*), založená na náhodném začlenění DNA krátkých, kolem 10 bp dlouhých primerů na templátovou jednořetězcovou DNA (*single-strand*, ssDNA) a jejich amplifikaci pomocí polymerázové řetězové reakce (*polymerase chain reaction*, PCR). Díky své malé délce bude použitý primer komplementární sekvencím na mnoha místech templátové DNA, takže nakonec na elektroforetickém gelu vidíme komplexní soubor proužků, který může být druhově specifický. Výhodou RAPD je to, že nemusíme mít žádnou předběžnou znalost o daném genomu. Bohužel, tato metoda trpí nízkou opakovatelností (např. i při stejných podmínkách v téže laboratoři můžeme získat odlišné výsledky), navíc často se mohou ve výsledné směsi vyskytnout i amplifikované fragmenty o sekvenci, která se nevyskytuje v DNA templátu.

Mnohem pokročilejší amplifikační metodou, která nevyžaduje žádnou znalost zkoumané sekvence, je výše zmíněná **AFLP** (*amplified fragment length polymorphisms*, polymorfismy délek amplifikovaných fragmentů). Výhodou této metody je, že dokáže rychle vygenerovat velké množství polymorfních markerů i u druhů, u kterých nemáme žádnou předchozí genetickou informaci. Na rozdíl od RAPD tato metoda využívá štěpení dvěma restrikčními enzymy a vyniká i vyšší mírou opakovatelnosti. Rozpoznávací místo jedné z restriktáz je kratší (např. enzym *Mse*I rozeznává sekvenci TTAA), kdežto druhá restriktáza vyhledává sekvenci delší a štěpí tedy méně často (např. *Eco*RI rozeznává sekvenci GAATTC). Tímto způsobem vzniká dostatek fragmentů, které budou na jednom konci

odstřiženy jiným enzymem než na konci druhém. Ve stejném kroku jsou na konce fragmentů enzymem ligázou připojeny krátké úseky DNA o známé sekvenci (adaptory), které se vážou na konce fragmentů, vytvořených příslušnou restrikční endonukleázou. Sekvence adaptorů jsou zvoleny tak, že je restrikční enzymy od fragmentů neodštěpí. Díky adaptorům získáme fragmenty, které mají na koncích známé sekvence, takže je lze amplifikovat pomocí PCR. Tímto způsobem vzniká velký počet fragmentů (jejich množství se zpravidla redukuje pomocí dvou běhů PCR), které jsou nakonec analyzovány pomocí automatického sekvenátoru.

Protože substituce i jediného nukleotidu většinou vedou ke změně sekundární struktury molekuly DNA, dá se tato skutečnost využít k analýze konformačního polymorfismu jednořetězcové DNA neboli SSCP (*single-strand conformational polymorphism*). Tato metoda spočívá v PCR se značenými primery nebo nukleotidy a následné denaturaci amplifikovaných produktů přidáním formamidu a NaOH a zahřátím. Potom jsou vzorky ssDNA rychle zchlazeny a podrobeny elektroforéze.

Kromě pořadí genů a SINE/LINE sekvencí můžeme využít i další typy tzv. vzácných genomických změn (*rare genomic changes*), například inzerční/deleční události, charakteristické sekvence (*signature sequences*), varianty genetického kódu mitochondriální a jaderné DNA, chromozomové přestavby, přestavby genů v mitochondriálním a chloroplastovém genomu nebo duplikace genů.

Jiné typy jinak velmi rozšířených molekulárních dat jsou pro fylogenetickou analýzu méně vhodné. Například tzv. mikrosatelity, tj. krátké, tandemově se opakující jednoduché sekvenční motivy zpravidla o 2–6 bp, jsou charakteristické vysokým mutačním tempem (u savců včetně člověka řádově $10^{-3}$ až $10^{-4}$, u *E. coli* dokonce $10^{-2}$ na jednu pozici na generaci), a proto příliš proměnlivé na to, aby poskytly spolehlivý obraz fylogenetických vztahů mezi druhy. Pouze v případě použití velmi málo variabilních mikrosatelitových lokusů bychom je mohli použít k analýze blízce příbuzných druhů.

## DATABÁZE SEKVENCÍ

### Databáze nukleotidových sekvencí

Existují tři světové veřejně dostupné nukleotidové databáze, paralelně udržované v rámci International Nucleotide Sequence Database Collaboration ve Spojených státech, Evropě a Japonsku. Jsou denně aktualizovány a vzájemně propojeny, takže zanesení sekvence do jedné z nich současně znamená její začlenění do ostatních databází. Jsou to:

**GenBank**, provozovaná v NCBI (National Center for Biotechnology Information, Bethesda, Maryland, USA) a dostupná na adrese http://www.ncbi.nlm.nih.gov/genbank/. Kompletní vydání je dostupné na FTP serveru NCBI a je vyhotovováno každé dva měsíce.

**EMBL** (European Molecular Biology Laboratory), kterou provozuje EMBL-EBI (European Bioinformatics Institute, Hinxton, Velká Británie). V současnosti tvoří součást databáze European Nucleotide Archive s adresou http://www.ebi.ac.uk/ena/.

**DDBJ** (DNA Data Bank of Japan), udržovaná NIG/CIB (National Institute of Genetics, Mishima, Japonsko) na http://www.ddbj.nig.ac.jp/.

Zpočátku byly tyto databáze obnovovány kurátory, kteří procházeli dostupnou literaturu, nyní jsou do nich nové údaje vnášeny přímo autory pomocí webových nástrojů.

## Hledání v databázové dokumentaci

Každá databáze musí být pochopitelně obsluhována určitým specializovaným softwarem. Typický obslužný program umožňuje uložit informaci o struktuře databáze, kurátorovi dovoluje vkládat a modifikovat jednotlivé položky, uživatelům umožňuje vyhledávat, prohlížet a stahovat požadované údaje. K tomu není nutné, aby uživatel vlastnil stejný obslužný program – jednak jsou tyto softwarové balíky poměrně drahé a jednak bohužel provozovatelé různých serverů téže databanky často volí odlišné programy. Proto jsou databáze zpravidla kopírovány a distribuovány ve formě jednoduchého textu (*plain text*, *flat file*) kódovaného ve formátu ASCII (American Standard Code for Information Interchange). Nejpoužívanějšími programovými balíky jsou Sybase a ORACLE.

V databázi můžeme vyhledávat buď zadáním určitého slova (přesněji řečeno sekvence znaků), nebo pomocí indexu vytvořeného ze slov v dokumentaci. V prvním případě postupujeme stejně, jako bychom prohledávali celou knihu, abychom nalezli určité slovo. V případě databází sekvencí však zpravidla nehledáme sekvence samotné, ale informace s nimi spojené. Přesto vyhledávání tímto způsobem většinou trvá poměrně dlouho. Nevýhodou je také riziko překlepu (např. „cytochrme“ místo „cytochrome“ – to je ovšem problém jakéhokoli vyhledávání zadáním hledaného slova), navíc jednotlivé databanky nemusí být jednotné v psaní některých pojmů, lze např. psát „HIV-1“ i „HIV1“. Proto je vhodné hledání několikrát opakovat pomocí různých výrazů nebo modifikace téhož výrazu.

Vyhledávání pomocí indexu je mnohem rychlejší než hledání slov (sekvence znaků). K nevýhodě předchozího způsobu vyhledávání, tj. že můžeme nalézt jen taková slova (resp. takové tvary slov), která tam byla předtím v témže tvaru vložena, ovšem musíme v tomto případě připočíst i možná rizika spojená s automatickým výběrem a ukládáním pojmů do indexu. Indexy jsou vytvářeny různými softwarovými systémy. Nejznámějším je SRS (Sequence Retrieval System), vyvinutý v EMBL-EBI. V indexu se většinou vyskytují jednotlivá slova, nemá proto smysl zadávat spojení dvou či více slov; lze je však spojit pomocí určitých speciálních znaků (z internetových prohlížečů známe např. booleanovské znaky „AND“, „OR“ atd., v systému SRS se používá znak „&“, např. „reverse & transcriptase“).

## Databáze GenBank

Bezpochyby nejznámější a nejpoužívanější databází je GenBank. Původně byla vytvořena a udržována v Los Alamos National Laboratory (LANL), ale počátkem 90. let byla z rozhodnutí Kongresu Spojených států amerických převedena pod NCBI. Jeho pracovníci nejprve procházeli dostupnou literaturu, nalezené sekvence manuálně vkládali do databáze a na základě informací získaných z příslušného publikovaného článku je anotovali, tj. přidávali relevantní poznámky. Jak už bylo uvedeno, dnes je tato praxe výjimečná, v drtivé většině případů jsou sekvence vkládány přímo autory sekvencí. Za touto změnou částečně stojí tlak většiny vydavatelů vědeckých periodik, kteří vyžadují, aby nukleotidové sekvence byly nejprve vloženy do veřejně dostupné databáze a byly tak čtenářům k dispozici už v okamžiku publikace článku. V současné době NCBI získává a procesuje kolem

20 000 přímých podání a 200 000 hromadných podání od velkých sekvenčních center (*high-throughput genomic sequences*) denně. Databáze tak exponenciálně narůstá, přičemž její rozsah se každých 10 měsíců zdvojnásobí a dá se očekávat, že toto tempo nadále poroste.

Databáze obsahuje údaje jak o DNA, tak RNA, ale podle existující konvence jsou i sekvence ribonukleových kyselin uvedeny s thyminem (T) místo uracilu (U). Každá sekvence v databázi obsahuje následující základní údaje:

- Název položky, jméno lokusu neboli identifikátor (ID). Každá sekvence má jediný ID, ten se ovšem může z verze na verzi měnit. Původně byla snaha zavádět mnemotechnické ID, s rostoucím počtem sekvencí však bylo od této praxe upuštěno.
- Přístupové číslo (*accession number*, AC). Toto číslo je v databázi jedinečné, ale jedna sekvence může mít více než jedno AC – k tomu dochází tehdy, když dvě nebo více sekvencí je zkombinováno do jedné. Přístupové číslo je velmi užitečným nástrojem pro vyhledávání sekvencí, protože zůstává mezi jednotlivými verzemi totožné a navíc existuje dohoda mezi výše zmíněnými databankami o tom, že stejným sekvencím jsou přiřazována stejná přístupová čísla.
- Číslo verze (*version*). Je odvozeno z AC a lze podle něho dohledat určitou sekvenci i poté, co bylo změněno její přístupové číslo. Bohužel, číslo verze bylo zavedeno až roku 1999, starší změny proto již vystopovat nelze.

Kromě těchto základních údajů je k sekvenci zpravidla připojena řada dalších informací o jejím zdroji, taxonomickém zařazení příslušného organismu atd. Sekvence je prezentována ve vlastním formátu po 60 bázích na řádek, přičemž každý řádek začíná pořadovým číslem první báze. Kliknutím na příkaz „FASTA“ ale můžeme formát GenBank změnit na formát FASTA (viz dále). Příkazem „Graphics“ můžeme získat další, graficky zpracovanou informaci o sekvenci.

Přímá podání (*submissions*) do databáze GenBank jsou zprostředkována buď přes webový formulář BankIt, nebo samostatným programem Sequin. BankIt je vhodný pro podání menšího počtu sekvencí s minimálními anotacemi. Obsahuje rozsáhlou nápovědu s konkrétními příklady. Sequin je vhodnější pro komplikovanější podání s mnoha sekvencemi nebo rozsáhlými anotacemi a lze ho stáhnout z FTP serveru NCBI. Jestliže chceme vložit větší počet příbuzných sekvencí, program akceptuje výstupy mnoha běžně rozšířených programů na seřazování sekvencí. Kompletní podání připravená pomocí Sequin jsou zaslána e-mailem na adresu gb-sub@ncbi.nlm.nih.gov, větší soubory mohou být podány prostřednictvím nástroje SequinMacrosend.

Prvním krokem po odeslání sekvencí do databáze je tzv. třídění neboli stanovení priorit (*triage*). Během této procedury, která trvá do 48 hodin od podání, personál databáze zjišťuje, zda soubor splňuje minimální kritéria stanovená pro zařazení do databáze a poté každé sekvenci přiřadí přístupové číslo. Všechny sekvence musí být alespoň 50 bp dlouhé. Akceptovány nejsou sekvence vytvořené *in silico* (tj. v počítači), diskontinuální sekvence obsahující vnitřní, nesekvenované mezerníky (*spacers*) nebo sekvence, pro něž neexistuje fyzický protějšek (např. sekvence směsi genomické DNA a mRNA). Mimoto je kontrolováno, zda jde o nové sekvence, nebo pouze o aktualizace už publikovaných sekvencí. Jakmile sekvence obdrží přístupové číslo, čeká je další kolo rozsáhlejšího procesování, např. kontrola biologické validity (kontrola sekvence aminokyselin pro kódující sekvence, taxonomická kontrola, porovnání se stávajícími záznamy), vektorové kontaminace, publikačního statutu a další formátování. Zkompletované položky jsou zaslány podavateli

ke kontrole a potom zveřejněny. K tomu dojde po pěti dnech, kdy lze ještě udělat poslední změny; podavatel ovšem může požádat o zvěřejnění až k pozdějšímu datu (např. až po oficiálním publikování článku, který s těmito sekvencemi pracuje).

Chceme-li z databáze GenBank získat sekvenci DNA, do horního okna domovské stránky vepíšeme přímo přístupové číslo požadované sekvence, které jsme získali například z publikovaného článku. Jestliže toto číslo nemáme k dispozici, napíšeme název taxonu a sekvence (genu), například „apodemus flavicollis rag1 gene“. V obou případech se ujistíme, že v rozbalovací roletě vlevo je zvoleno „Nucleotide“. Zde je příklad sekvence kontrolní oblasti mtDNA a částí okolních genů:

**Mus macedonicus spretoides isolate 16654 tRNA-Thr gene, partial sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gene, complete sequence; mitochondrial**

GenBank: EU106188.1

FASTA Graphics PopSet

```
LOCUS       EU106188               1063 bp    DNA     linear   ROD 16-NOV-2007
DEFINITION  Mus macedonicus spretoides isolate 16654 tRNA-Thr gene, partial
            sequence; and tRNA-Pro gene, D-loop, and tRNA-Phe gene, complete
            sequence; mitochondrial.
ACCESSION   EU106188
VERSION     EU106188.1  GI:157265957
KEYWORDS    .
SOURCE      mitochondrion Mus macedonicus spretoides
  ORGANISM  Mus macedonicus spretoides
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.
REFERENCE   1  (bases 1 to 1063)
  AUTHORS   Macholan,M., Vyskocilova,M., Bonhomme,F., Krystufek,B., Orth,A. and
            Vohralik,V.
  TITLE     Genetic variation and phylogeography of free-living mouse species
            (genus Mus) in the Balkans and the Middle East
  JOURNAL   Mol. Ecol. 16 (22), 4774-4788 (2007)
   PUBMED   17908218
REFERENCE   2  (bases 1 to 1063)
  AUTHORS   Macholan,M., Vyskocilova,M., Bonhomme,F., Krystufek,B., Orth,A. and
            Vohralik,V.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-AUG-2007) Laboratory of Mammalian Evolutionary
            Genetics, Institute of Animal Physiology and Genetics, Acad. Sci.
            Czech Rep., Veveri 97, Brno CZ-60200, Czech Republic
FEATURES             Location/Qualifiers
     source          1..1063
```

```
                     /organism="Mus macedonicus spretoides"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /isolate="16654"
                     /sub_species="spretoides"
                     /db_xref="taxon:270352"
                     /country="Israel: Dor"
                     /note="type: ISR1"
     tRNA            <1..37
                     /product="tRNA-Thr"
     tRNA            38..105
                     /product="tRNA-Pro"
     D-loop          106..983
     gap             477..588
                     /estimated_length=112
     tRNA            984..1050
                     /product="tRNA-Phe"
ORIGIN
        1 tgtaaacctg aaatgaagat cttctcttct caaggcatca agaagaagga actttttccc
       61 caccgccaac acccaaagct ggtattctaa ttaaactact tcttgagtac ataaatttac
      121 atagtacaat agtacattta tgtatatcgt acattaaatt ataatcccca agcatataag
      181 caagtaaatt aaattaatta tataacacat aaaattaata ctcaacataa tatgtcatac
      241 accatgaata ttacaccaag tacattaaat taatgtttta aagacatatc tgtgttatct
      301 gacatacacc ataaagtcat aaacccttct cttccatatg actatcccct tccccatttg
      361 gtctattaat ctaccatcct ccgtgaaacc aacaacccgc ccacctatgc ccctcttctc
      421 gctccgggcc cattaaactt gggggtagct aaactgaaac tttatcagac atctgg
          [gap 112 bp]    Expand Ns
      589                                                       ca tttggtattt
      601 ttttattttg gcctactttc atcaacatag ccgtcaaggc atgaaaggac agcacccagt
      661 ctagacgcac ctacggtgaa gaatcattag tcctcataac ccaatcaccc aaggctaatt
      721 attcatgctt gttagacata aaattattca ataccaggtt ttaactcttc aaaccccccc
      781 tcacccccac cctcttaatg ccaaacccca aaaacattaa gaacttgaag acatatatta
      841 ttaactatca aaccctatgt cctgatcaat tctagtagtt caaaaaatat gacttatatt
      901 ttagttcttg taaaaatttt gcaaaattat gccccataaa ccaaaactct tattacaccc
      961 tattacgcaa taaataacgg taggttaatg tagcttaata aaaagcaaag cactgaaaat
     1021 gcttagatgg ataattatat cccataaaca caaaggtttg gtc
//
```

Spektrum informací, které můžeme získat z databáze GenBank, je ovšem mnohem širší, například o konkrétních genech či celých genomech, EST (Expressed Tag Sequences), GSS (Genome Survey Sequences), SNP, proteinových sekvencích, proteinové struktuře atd. Všechny možnosti jsou uvedeny v seznamu, který nalezneme v rozbalovací roletě vlevo nahoře.

## Program BLAST

Jakmile získáme neznámou sekvenci, zpravidla chceme zjistit, kterému organismu patří, a i když zdroj sekvence známe, měli bychom si ověřit, že jsme neosekvenovali jinou DNA, kterou mohl být náš materiál kontaminován. V neposlední řadě můžeme chtít nalézt sekvence příbuzných organismů. K tomuto účelu slouží program BLAST (Basic Local Alignment Search Tool), který najdeme na adrese http://blast.ncbi.nlm.nih.gov/Blast.cgi. Algoritmus tohoto programu nejprve rozloží zadanou sekvenci na kratší fragmenty neboli „slova" (*words*) a potom hledá shodu mezi těmito fragmenty a sekvencemi v databázi. Jakmile je nalezena část nějaké sekvence shodná nebo velmi podobná hledanému fragmentu, hledání se od tohoto úseku rozšiřuje v obou směrech. Míra shody mezi sekvencemi je kvantifikována pomocí skóre a sekvence s největší shodou jsou uvedeny v tabulce od nejvyšších po nejnižší skóre. BLAST umožňuje hledání v nukleotidových i proteinových databázích. Obsahuje i rozsáhlé genomové databáze nejen pro modelové organismy, jako jsou člověk (Human), myš (Mouse), potkan (Rat), šimpanz (*Pan troglodytes*), tur domácí (*Bos taurus*), kur domácí (*Gallus gallus*), danio pruhované (*Danio rerio*), včela (*Apis mellifera*), octomilka (*Drosophila melanogaster*), huseníček (*Arabidopsis thaliana*), rýže (*Oryza sativa*) a mikroorganismy (Microbes), ale pro celou řadu dalších.

K nalezení nukleotidové sekvence můžeme použít jeden ze tří programů: BLASTN, MEGABLAST nebo DISCONTIGUOUS MEGABLAST. MEGABLAST je speciálně uzpůsoben k efektivnímu vyhledávání dlouhých, velmi podobných seřazených sekvencí, a proto je zvlášť vhodný pro nalezení identické shody s naší dotazovanou sekvencí (*query*). Jestliže chceme vyhledat podobné sekvence jiných organismů, je vhodnější program BLASTN, protože sekvenci rozděluje na kratší „slova". Citlivost hledání může být zvýšena jejich zkrácením z přednastavené hodnoty až na 7. Senzitivnější vyhledávání může být dosaženo i použitím programu DISCONTIGUOUS MEGABLAST. Jeho algoritmus místo toho, aby vyžadoval naprostou shodu „slov", odkud by pokračoval v hledání, používá nesouvislá „slova". Například v kódujících sekvencích bere v úvahu degenerovanost genetického kódu, a proto hledá shodu vždy jen v první a druhé pozici kodonu, kdežto rozdíly ve třetí pozici ignoruje. Tímto způsobem DISCONTIGUOUS MEGABLAST dosahuje při stejné délce „slova" vyšší citlivosti hledání než BLASTN. DISCONTIGUOUS MEGABLAST lze ovšem použít i pro nekódující sekvence.

Hledání shody mezi nukleotidovými sekvencemi však není příliš vhodné pro nalezení homologních protein kódujících oblastí u jiných organismů. K tomu slouží vyhledávání v proteinových nebo translatovaných databázích. V prvním případě jsou k dispozici programy BLASTP, PSI-BLAST (Position-Specific Iterated BLAST), PHI-BLAST (Pattern-Hit Initiated BLAST) a DELTA-BLAST, ve druhém programy BLASTX (prochází proteinovou databázi se zadanou translatovanou nukleotidovou sekvencí), TBLASTN (prochází translatovanou nukleotidovou databázi se zadanou proteinovou sekvencí) a TBLASTX (prochází translatovanou nukleotidovou databázi se zadanou translatovanou nukleotidovou sekvencí). Kromě uvedených základních programů BLAST obsahuje i řadu specializovaných typů vyhledávání.

Jako příklad si můžeme ukázat vyhledání sekvence cytochromu *b* gibona lar (*Hylobates lar*). Byla získána z databáze GenBank zadáním požadavku „hylobates lar cytochrome b"; vybrána byla sekvence Y13301 o délce 1141 bp, do databáze zadaná 18. 4. 2005 a publikovaná v článku Hall et al. (1998) „Evolution of the *Gibbon* subgenera inferred from cytochrome *b* DNA sequence data" (*Molecular Phylogenetics and Evolution*, 10: 281–286).

Otevřeme program BLAST a v sekci „Basic BLAST“ vybereme možnost „nucleotide blast“. Do horního okna zkopírujeme sekvenci ve formátu FASTA (viz str. 28). V okně „Job Title“ se objeví název sekvence: „gi|2765309|emb|Y13301.1| Hylobates lar mitochondrial...“. Všechny další vstupní informace ponecháme tak, jak jsou přednastaveny (dole v oddíle „Program Selection“ je zaškrtnuta volba „Highly similar sequences (megablast)“), a klikneme na tlačítko „BLAST“. Ve výsledné tabulce se po souhrnné informaci o zadané sekvenci objeví grafické znázornění výsledků nazvané „Distribution of 100 Blast Hits on the Query Sequence“. Protože všech 100 sekvencí se s dotazovanou sekvencí shoduje ve více než 200 bázích, jsou všechny znázorněny červenou barvou. Pod grafikou následuje seznam sekvencí s významnou shodou („Sequences producing significant alignments“). Na prvním místě je sekvence, kterou jsme zadali a která má pochopitelně 100% shodu. Míra shody mezi sekvencemi a další informace jsou uvedeny ve sloupcích vpravo („Max score“, „Total score“, „Query cover“, „E value“, „Ident“, „Accession“). Čím je celkové skóre vyšší, tím je vybraná sekvence příbuznější sekvenci zadané. Hodnota E („E value“) udává pravděpodobnost, že nalezená shoda mezi sekvencemi je pouze náhodná. Konečně míra identity („Ident“) je procentuálním vyjádřením podílu shodných pozic a jejich celkového počtu. Například na 58. řádku tabulky nalezneme sekvenci „Hylobates agilis mitochondrial cytb gene“ s přístupovým číslem AJ010583.1. Kliknutím na název sekvence dostaneme detailní přehled o shodě. V našem případě je přiřazená sekvence dlouhá 1139 bp (protože námi zadaná sekvence byla dlouhá 1141 bp, je „Query cover“ = 99 %) a mezi oběma sekvencemi bylo nalezeno 1070 shodných pozic a žádné mezery („Identities“ = 1070/1139 (94 %); „Gaps“ = 0/1139 (90 %)). Následuje grafické znázornění seřazení obou sekvencí:

| Score | Expect | Identities | Gaps | Strand |
|---|---|---|---|---|
| 1722 bits(932) | 0.0 | 1070/1139(94%) | 0/1139(0%) | Plus/Plus |

```
Query  1     ATGACCCCCCTGCGCAAAACTAATCCACTAATAAAACTAATCAACCACTCACTTATCGAC  60
             |||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     ATGACCCCCCTACGCAAAACTAATCCACTAATAAAACTAATCAACCACTCACTTATCGAC  60

Query  61    CTTCCAGCCCCATCCAACATTTCTATATGATGAAACTTTGGTTCACTCCTAGGCGCCTGC  120
             ||||| ||||||||||||||| || |||||||||||||||| ||||||||||||||||||
Sbjct  61    CTTCCGGCCCCATCCAACATCTCCATATGATGAAACTTTGGCTCACTCCTAGGCGCCTGC  120

Query  121   CTGATCCTCCAGATCATCACAGGATTATTTTTAGCCATACACTACACACCAGATGCCTCC  180
             || |||||||| || ||||||||||||||||||||||||||||||||||| ||||||
Sbjct  121   CTAATCCTCCAAATTATCACAGGATTATTTTTAGCCATACACTACACACCAGACGCCTCC  180

Query  181   ACAGCTTTCTCATCAGTAGCTCACATCACCCGAGACGTAAACTACGGCTGAATCATCCGC  240
             |||||||||||||||||| || |||||||||||||||||||||||||| ||||||
Sbjct  181   ACAGCTTTCTCATCAGTAGCCCATATCACCCGAGACGTAAACTACGGCTGAATTATCCGC  240
```

Kliknutím na přístupové číslo nalezené sekvence se dostaneme na příslušnou stránku sekvence cytochromu *b* gibona tmavorukého (*H. agilis*) v databázi GenBank:

```
LOCUS       AJ010583               1141 bp    DNA     linear   PRI 15-APR-2005
DEFINITION  Hylobates agilis mitochondrial cytb gene.
ACCESSION   AJ010583
VERSION     AJ010583.1  GI:4160515
KEYWORDS    cytb gene; cytochrome b.
SOURCE      mitochondrion Hylobates agilis (agile gibbon)
  ORGANISM  Hylobates agilis
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hylobatidae; Hylobates.

[....]

ORIGIN
        1 atgacccccc tacgcaaaac taatccacta ataaaactaa tcaaccactc acttatcgac
       61 cttccggccc catccaacat ctccatatga tgaaactttg gctcactcct aggcgcctgc
[....]
     1021 cagccggtaa gctacccgtt tatcaccatt ggacaaatgg catccgtact gtacttcacc
     1081 acaatcctag tactaatgcc agccgcctcc ctagtcgaaa acaaaatact caaatgaact
     1141 t

//
```

## Formáty souborů

Jednotlivé fylogenetické programové balíky bohužel zpravidla využívají odlišné formáty vstupních a výstupních souborů. Ruční editování dlouhých sekvencí je pochopitelně značně nepraktické. Naštěstí pro uživatele existuje několik programů, které převádějí soubory z jednoho formátu do druhého. Mezi nejznámější a nejvíce rozšířené formáty, se kterými se můžeme setkat při fylogenetické analýze, patří FASTA, Clustal, NEXUS a PHYLIP. FASTA je nejjednodušší formát s jedinou nepřerušenou sekvencí. Začíná vždy znakem „>“, za kterým následuje (bez mezery) tzv. identifikátor, který může být následován popisem sekvence. Identifikátor i popis nejsou povinné. Na dalším řádku začíná vlastní sekvence. Doporučuje se, aby všechny řádky textu (tj. první popisný řádek i vlastní sekvence) nepřesahovaly 80 znaků. Příklad šesti krátkých sekvencí ve formátu FASTA:

```
>H_sapiens [= identifikátor] 350 bp cytochrom b [= popis]
ATGACCCCAATACGCAAAATTAACCCCCTAATAAAATTAATTAACCACTCATTCATCGACCTCCCCACCC
CATCCAACATCTCCGCATGATGAAACTTCGGCTCACTCCTTGGCGCCTGCCTGATCCTCCAAATCACCAC
AGGACTATTCCTAGCCATACACTACTCACCAGACGCCTCAACCGCCTTTTCATCAATCGCCCACATCACT
CGAGACGTAAATTATGGCTGAATCATCCGCTACCTTCACGCCAATGGCGCCTCAATATTCTTTATCTGCC
TCTTCCTACACATCGGGCGAGGCCTATATTACGGATCATTTCTCTACTCAGAAACCTGAAACATCGGCAT
```