Česká zemědělská univerzita v Praze

Fakulta agrobiologie, potravinových a přírodních zdrojů

Katedra zoologie a rybářství

Využití metod sekvenace DNA v ichtyologii

Bakalářská práce

Autor práce: Jakub Friedl

Vedoucí práce: doc. Ing. Lukáš Kalous, Ph.D.

Čestné prohlášení

Prohlašuji, že svou bakalářskou práci "Využití metod sekvenace DNA v ichtyologii" jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu literatury na konci práce. Jako autor uvedené bakalářské práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 20. 4. 2015 _____

Využití metod sekvenace DNA v ichtyologii

Souhrn

Několik prvních stránek práce je věnováno základním informacím o deoxyribonukleové kyselině (DNA), historii jejího objevu a jejím nejdůležitějším vlastnostem.

Další kapitola se věnuje principům sekvenačních metod, od historických počátků krátce po objevení struktury DNA, kdy bylo sekvenování DNA považováno za obtížnější problém než sekvenování proteinů, přes první efektivní sekvenační metody včetně metody Sangerovy, až po masivně paralelní sekvenační metody druhé generace a nejnovější metody schopné sekvenovat jednotlivé molekuly DNA. Součástí je i shrnutí vlastností, výhod a nevýhod dnes nejpoužívanějších sekvenačních metod druhé generace.

Následuje část, která se zabývá zpracováním získaných sekvenčních dat, určováním bází a skládáním sekvencí, alignmentem a databázemi sekvencí a jejich vyhledáváním.

Dále se práce zabývá využitím sekvenování DNA (obvykle části mitochondriálního genu pro podjednotku I cytochrom c oxidázy) k identifikaci známých druhů (technika DNA barcodingu). Jsou uvedeny příklady jeho využití pro identifikaci rybích produktů na trhu (kde je velmi vysoký výskyt různých podvodů včetně zdraví ohrožujících), identifikaci rybí potravy z morfologicky již neurčitelného obsahu trávící soustavy, využití z hlediska biologické bezpečnosti (invazivní druhy, přenašeči chorob) a ochrany druhů při importu ryb přes hranice až po využití při identifikaci druhů ze vzorků prostředí (metabarcoding).

Následující část se věnuje problematice fylogenetické analýzy na základě sekvence DNA. Popisuje základní metody fylogenetické analýzy, zejména principy metody maximální úspornosti, maximální věrohodnosti i bayesovské analýzy. V další části se věnuji nejznámějším programům pro fylogenetickou analýzu. Poslední část rešerše se věnuje využití masivně paralelního sekvenování a mnoha lokusů ve fylogenetické analýze. V závěru autor shrnuje současný rozvoj sekvenačních metod a jejich využití a zamýšlí se nad jejich rostoucím významem.

Klíčová slova: fylogenetika, program, sekvenace, analýza, barcoding

The use of DNA sequencing methods in ichthyology

Summary

The first few pages of the thesis are dedicated to the basic information about the deoxyribonucleic acid (DNA), the history of its discovery and its most important properties.

The next chapter focuses on the principles of the DNA sequencing methods, from the very beginnings short after the discovery of the DNA structure, to the first effective sequencing methods including the Sanger method, the massively parallel sequencing methods of the second generation and the most recent single molecule sequencing methods. The summary of the properties, advantages and disadvantages of the most popular sequencing methods currently in use is included.

The next part of the thesis describes the processing of raw sequencing data, including the base-calling, the sequence assembly, alignment, sequence databases and searching.

DNA barcoding, the method which uses DNA sequencing (usually sequencing of a part of the subunit I of the mitochondrially encoded cytochrome c oxidase) for species identification, is described in the next part. Examples of the use of this method are given, such as the identification of mislabelled fish food products (frauds in this industry are common and some may be harmful to health), the identification of morphologically unrecognizable prey in fish digestive systems, the use in biosecurity control and biodiversity protection, and the use for identification of species from environmental samples (metabarcoding).

The next chapter is dedicated to the phylogenetic analysis based on DNA, especially the DNA sequence. The basic methods are described, including the principles the maximum parsimony, the maximum likelihood and the Bayesian inference. The most well know programs for phylogenetic analysis are described. The last part focuses on the use of modern masivelly paralel sequencing methods and multiple locuses in the phylogenetical analysis. In the end, the author summarizes the current development and considers the increasing importance of the DNA sequencing in biology and ichthyology.

Keywords: phylogenetics, program, sequencing, analysis, barcoding

Obsah

1 Úvod

Velký rozvoj molekulární biologie a genetiky začal objevem struktury deoxyribonukleové kyseliny (DNA) v padesátých letech minulého století, na kterém se podíleli Francis Crick, James Watson, Maurice Wilkins a Rosalind Franklin. Postupně byly vyvíjeny a zdokonalovány metody čtení, tedy sekvenování, informace uložené v DNA. V sedmdesátých letech byly vyvinuty první skutečně efektivní sekvenovací technologie, zejména Sangerova dideoxy metoda. Ta byla využita k osekvenování lidského genomu. Projekt lidského genomu, zahájený v roce 1990, trval více než deset let a znamenal další významné zdokonalení sekvenovacích technik. Již druhý osekvenovaný obratlovčí genom však byl rybí (čtverzubec fugu). Sekvenování DNA byla v té době stále velmi drahá a zdlouhavá záležitost a možnosti využití proto byly omezené.

V roce 2005 ale začal rozvoj masivně paralelních sekvenovacích technik, které, narozdíl od Sangerovy dideoxy metody, umožňují sekvenovat velké množství fragmentů DNA paralelně a tím sekvenování významným způsobem zrychlují a zlevňují. Nejnovější sekvenační technologie pak umožňují sekvenovat jednotlivé molekuly DNA, což nadále rozšiřuje možnosti využití sekvenování.

Samotné fyzické čtení sekvence ale není jedinou důležitou částí procesu, neméně důležité je následné zpracování a analýza sekvence, což je se stále větším přísunem sekvenčních dat stále náročnějším úkolem. Sekvenování DNA nalézá stále větší uplatnění i ve výzkumu ryb, ichtyologii. Klasickým využitím je výzkum fylogeneze (a biogeografie) ryb, kde spolu s masivně paralelním sekvenováním dochází k posunu od fylogenií postavených na jednom nebo několika málo genech k fylogeniím postavených na větších částech genomu, v budoucnosti možná i na kompletních genomech.

Významným využitím sekvenace DNA je také identifikace známých druhů (metoda tzv. DNA barcodingu), a to dnes vzhledem k pokroku sekvenovacích metod již nejen ze vzorků tkáně, ale i z tzv. environmentální DNA, tedy například vzorku vody, ve které plavaly ryby. DNA

barcoding má význam v mnoha oblastech ichtyologie, od kontroly pravosti potravin z ryb, přes identifikaci druhů importovaných přes hranice (ochrana biodiverzity, invazivní druhy), přes možnost identifikace obsahu trávicí soustavy ryb v ekologických studiích, možnost identifikace různých vývojových stádií, morfologicky obtížně rozlišitelných a v mnoha dalších situacích.

2 Cíl práce

Cílem práce je podat ucelený obraz o vývoji metod sekvenace DNA, zejména o rychlém až překotném vývoji v posledních deseti letech, který zpřístupňuje sekvenování DNA pro stále širší okruh využití, a o zpracování a využití získaných dat v různých oblastech ichtyologie, jako je fylogenetická analýza, která je důležitým aspektem studia biologie ryb, nebo identifikace rybích druhů, která je důležitá v mnoha oborech lidské činnosti.

3 Literární rešerše

3.1 Deoxyribonukleová kyselina

Deoxyribonukleovou kyselinu (DNA) objevil v roce 1869 švýcarský lékař Friedrich Miescher. Při biochemické analýze hnisu z nemocničních obvazů objevil dosud neznámou substanci bohatou na fosfor a odlišnou od proteinů, kterou nazval nuklein. V průběhu 19. století byla DNA chemicky analyzována (například německý biochemik Albrecht Kossel popsal dusíkaté báze), ale ještě dlouho po Miescherově smrti v roce 1895 vědci nepřipisovali nukleinu velký význam. Domnívali se, že nositelem dědičné informace musí být proteiny, protože jsou složitější a obsahují dvacet různých aminokyselin. Pouhé čtyři různé dusíkaté báze nesoucí informaci v DNA se zdály být pro kódování obrovského množství genetické informace příliš málo. (Dahm, 2005).

Do středu všeobecného zájmu se DNA dostala až ve čtyřicátých letech 20. století, kdy bylo prokázáno, že DNA je nositelkou genetické informace (Dahm, 2005). Přelomový byl zejména známý Averyho-MacLeodův-McCartyho experiment, ve kterém autoři prokázali, že substancí způsobující transformaci baktérií je právě DNA (Avery et al., 1944). V roce 1953 James Watson, Francis Crick, Rosalind Franklinová a Maurice Wilkins analyzovali a podrobně popsali strukturu dvoušroubovice DNA. V šedesátých letech 20. století pak Robert Holley, Har Gobind Khorana, Marshall Nirenberg a další rozluštili genetický kód popisující způsob, jakým jsou v sekvenci DNA kódované proteiny (Dahm, 2005).

DNA je makromolekula, biopolymer. Základní jednotkou DNA je nukleosid - pětiuhlíkatý cukr deoxyribosa, na jejímž prvním uhlíku je N-glykosidickou vazbou navázán jeden ze čtyř druhů dusíkatých bází - v DNA se vyskytují purinové báze adenin (A) a guanin (G) a pyrimidinové báze cytosin (C) a thymin (T). Deoxyribózy jsou mezi pátým a třetím uhlíkem propojeny fosfodiesterovou vazbou (nukleosid spolu s fosfátovou skupinou se nazývá nukleotid) a tvoří tak orientovaný řetězec (má takzvaný 5' a 3' konec). Genetická informace je zapsána v pořadí dusíkatých bází v řetězci DNA (Flegr, 2009).

Důležitou vlastností dusíkatých bází je jejich komplementarita - vytváří komplementární páry spojené vodíkovými můstky. Za normálních okolností se v DNA páruje vždy purinová báze s pyrimidinovou, konkrétně adenin s thyminem (dva vodíkové můstky) a guanin s cytosinem (tři vodíkové můstky). Takto se nepárují sousední báze v jednom vlákně DNA, ale báze ze dvou různých vláken (případně vzdálené báze z jednoho vlákna pokud vytvoří smyčku). DNA se v organismu běžně vyskytuje ve dvouřetězcové formě, tvořené dvěma antiparalelními (5'-konec jednoho řetězce se páruje s 3'-koncem druhého řetězce) řetězci propojenými vodíkovými můstky mezi dvojicemi komplementárních dusíkatých bází (Flegr, 2009). Přestože je vodíkový můstek poměrně slabá vazba, množství vodíkových můstků v mnoha párech bází drží obě vlákna poměrně pevně u sebe.

Díky jednoznačnému přiřazení vzájemně si komplementárních bází obsahují oba řetězce dvouřetězcové DNA stejnou genetickou informaci. Toho je využito například v procesu replikace DNA, kdy se dvouřetězcová DNA dočasně rozdělí a oba jednotlivé řetězce slouží jako předloha k syntéze řetězce komplementárního. Tak z jedné dvouřetězcové molekuly DNA vzniknou molekuly dvě obsahující stejnou genetickou informaci (Flegr, 2009).

Genetická informace všech známých buněčných organismů i mnoha virů je zapsána v pořadí dusíkatých bází respektive nukleotidů v molekule DNA, tedy v její primární struktuře (zbývající viry a viroidy využívají jako nositelku genetické informace velmi podobnou kyselinu ribonukleovou - RNA). Sekvenci nukleotidů lze zjišťovat procesem sekvenování, který byl v minulosti nesmírně náročný a drahý, ale v současnosti se stává stále dostupnější a jednodušší technikou (Flegr, 2009).

Je ale nutné mít na zřeteli, že část dědičné informace určující vlastnosti organismu se nachází i mimo primární strukturu nukleové kyseliny. Tyto informace se nazývají epigenetické a patří mezi ně například modifikace DNA a na ni navázaných proteinů pomocí methylace, acylace či dalších skupin. Takovéto modifikace mohou pozitivně nebo negativně ovlivňovat aktivitu příslušného úseku DNA, přestože se primární struktura nemění. Po replikaci může být např. methylace na nově syntetizovaném řetězci doplněna podle řetězce původního specializovaným enzymem. Epigenetická informace může být uložena i v dalších částech

buňky (Flegr, 2009).

Důležitým pojmem s nejednoznačnou definicí je gen. Tradičně se gen definuje jako genetická informace ovlivňující nějakou rozpoznatelnou vlastnost jedince, výskyt určitého znaku nebo jeho konkrétní formu (Flegr, 2009). V molekulární biologii se obvykle termín gen používá ve významu cistronu, tedy souvislého úseku DNA kódujícího například určitou ribozomální RNA nebo protein (prostřednictvím mRNA, která může ještě býti upravována cis-sestřihem) (Flegr, 2009). Z pohledu evoluční biologie se může v podstatě každý nukleotid v regulačních a kódujících oblastech DNA chovat jako samostatný gen (Flegr, 2009).

Veškerá DNA nacházející se v buňce tvoří genom. U živočichů včetně všech ryb se DNA vyskytuje ve dvou částech buňky. Většina DNA se nachází v buněčném jádře v podobě chromozomů, velkých lineárních molekul DNA. To je jaderný genom. U typických diploidních biparentálních živočichů se každý jaderný chromozóm vyskytuje ve dvou potenciálně odlišných kopiích, jedné zděděné po otci, druhé po matce (Flegr, 2009).

3.1.1 Mitochondriální DNA

Menší část genetické informace se nachází v mitochondriích, buněčných organelách, které podle tzv. endosymbiotické teorie kdysi byly samostatnými endosymbiotickými organismy. To je mitochondriální genom (Flegr, 2009). Mitochondrií je v buňce obvykle mnoho, proto je mnoho i kopií mitochondriální DNA, obvykle jsou ale zvířata homoplasmická (všechny mitochondrie obsahují stejný genom), což je pravděpodobně způsobeno efektem úzkého hrdla při vývoji samičích pohlavních buněk. Pro analýzy je to výhoda, z tohoto pravidla ale existují výjimky (Avise, 2004). Mitochondriální DNA je poměrně malá, obvykle cirkulární kovalentně uzavřená dvouřetězcová molekula, o délce asi 15-20 kilobází (Avise, 2004). Například u dánia pruhovaného (*Danio rerio*) obsahuje 16596 nukleotidů a kóduje pouze 13 proteinů, 22 tRNA a 2 rRNA (Broughton et al, 2001). Sekvence živočišné mitochondriální DNA se oproti jaderné DNA vyvíjí velmi rychle, zejména její úsek zvaný D-smyčka. Proto je velmi vhodná pro analýzy příbuzných druhů nebo analýzy vnitrodruhové (Flegr, 2009), včetně nedávno vzniklé populační struktury (Avise, 2004). Většina mitochondriálního genomu má kodující funkci, pseudogeny, repetitivní DNA a podobně jsou vzácné nebo zcela chybí. Pořadí se naopak příliš nemění a proto lze využít v pro rozlišení vyšších taxonů (Avise, 2004).

Pro mitochondriální DNA je typická maternální dědičnost, mitochondrie a jejich genom jsou v naprosté většině případů děděny po matce. To omezuje rekombinaci mitochondriální DNA. Mitochondriální DNA je tak maternálním markerem, který je přenášen asexuálně i u jinak sexuálního druhu. Mitochodnriální genomy se tak vlastně z evolučního hlediska chovají jako velké supergeny a označují se jako haplotypy (Avise, 2004).

I z pravidla o maternální dědičnosti mitochondrií ale existují výjimky (Avise, 2004). U mlžů z čeledi slávkovití (Mytilidae) je známá tzv. dvojitě uniparentální dědičnost, kdy je většina mitochondrií děděna od matky, ale samci dědí mitochondrie od obou rodičů, přičemž otcovské se vyskytují v gonádách, zatímco zbytek těla obsahuje mateřské (Zouros, 2000). Z hlediska ichtyologie jsou zajímavé například zaznamenané případy u hybridních makrel rodu *Scomberomorus* (Morgan et al, 2013).

3.2 Metody sekvenování DNA

3.2.1 Počátky sekvenování DNA

V roce 1951 byla publikována sekvence prvního proteinu, B-řetězce bovinního inzulínu (Sanger a Tuppy, 1951). První kompletní sekvence nukleové kyseliny, 77 nukleotidů dlouhé kvasinkové alanin tRNA, byla publikována v roce 1965 (Holley et al., 1965). Sekvenace DNA byla ale náročnějším úkolem. Od objevu struktury DNA (Watson a Crick, 1953) do první částečné sekvenace krátkého úseku DNA, kohezních konců DNA bakteriofága λ (Wu a Kaiser, 1968), uplynulo 15 let, kompletní 12 nukleotidů dlouhá sekvence ale byla zveřejněna až o tři roky později v roce 1971 (Wu a Taylor, 1971).

Příčinou obtíží se sekvenováním DNA bylo několik – molekuly DNA jsou si navzájem chemicky velmi podobné a bylo proto obtížné získat větší množství homogenní DNA, řetězce DNA jsou výrazně delší než řetězce proteinů či malých RNA a především, v této době nebyly známé žádné enzymy, které by DNA štěpily specificky v místě určitých bází (Wu, 1970), zatímco při sekvenování proteinů a RNA se specifické proteinázy a ribonukleázy využívaly (Hutchison III, 2007).

Sekvence kohezních konců bakteriofága λ byla zjištěna velmi komplikovaným způsobem, zahrnujícím značení radioaktivním fosforem (^{32}P) a tritiem (^{3}H) v různých kombinacích, prodlužování kratšího řetězce DNA polymerázou a postupné skládání informací z řady různých experimentů (Wu a Taylor, 1971). Tato metoda byla použitelná jen pro koncové úseky DNA λ a podobných bakteriofágů, generalizaci umožnilo použití oligonukleotidů jako primerů pro sekvenační reakci, což Wu publikoval v roce 1971 (Hutchison III, 2007). Významným průlomem byl také objev restrikčních endonukleáz typu II v roce 1970. Tyto enzymy umožňují štěpit DNA v místě krátké specifické sekvence a brzy bylo nalezeno velké množství restrikčních endonukleáz specifických pro různé sekvence (Hutchison III, 2007).

Tyto rané metody ale nebyly dostatečně silné pro sekvenaci delších úseků DNA jako jsou kompletní genové sekvence. Bylo s jejich pomocí ale objeveno několik regulačních sekvencí, například operátor *lac* operonu *Escherichia coli* (Hutchison III, 2007). Až v druhé polovině sedmdesátých let byly vynalezeny metody, které umožnily skutečně efektivní sekvenování DNA. Walter Gilbert a Frederick Sanger za ně v roce 1980 společně obdrželi polovinu Nobelovy ceny na chemii (Nobelprize.org, 2014).

3.2.2 Sangerova „plus a minus“ metoda

Roku 1975 publikoval Frederick Sanger, autor sekvenace inzulínu, novou metodu sekvenace DNA. Stejně jako Wu využil DNA polymerázu prodlužující vlákno začínající primerem. Významnou novinkou bylo využití elektroforézy na polyakrylamidovém gelu pro separaci a roztřídění produktů syntézy (Hutchison III, 2007).

V první části sekvenace byly podmínky nastaveny tak, aby syntéza nového vlákna byla pomalá a asynchronní. Výsledkem měla být směs radioaktivně značených (^{32}P) vláken pokud možno všech možných délek. Po přečištění pokračovala sekvenace ve dvou systémech („plus“ a „minus“), každý z nich rozdělen do čtyř reakcí (Sanger a Coulson, 1975).

V reakčních směsích systému „minus“ chyběl vždy jeden ze čtyř nukleotidů (dNTP), zbývající tři byly přítomny. Použita byla DNA polymeráza I zbavená exonukleázové aktivity (Klenowův fragment). Prodlužování vláken probíhalo, dokud se nezastavilo kvůli

nedostupnému nukleotidu. Například v reakci s chybějícím dATP syntéza každého vlákna skončila na pozici před adeninem (Sanger a Coulson, 1975).

V reakčních směsích systému „plus“ byl vždy přítomen pouze jeden ze čtyř nukleotidů (dNTP). Použita byla T4 DNA polymeráza (z *Escherichia coli* infikované bakteriofágem T4). Ta má v přítomnosti pouze jednoho typu nukleotidu silnou 3’->5’-exonukleázovou aktivitu, která se zastaví, jakmile narazí na bázi odpovídající nukleotidu dostupnému v reakční směsi. Vlákna se proto zkracovala a výsledkem byla směs vláken končících dostupným nukleotidem (Sanger a Coulson, 1975).

Produkty všech osmi reakcí byly elektroforeticky rozděleny podle své délky na polyakrylamidovém gelu a zobrazeny autoradiograficky (gel byl přiložen na rentgenový filmový materiál) (Sanger a Coulson, 1975).

Teoreticky by oba systémy samostatně měly poskytnout kompletní sekvenci, podmínkou však je, že v počáteční reakci vzniknou skutečně vlákna všech možných délek. To však v praxi nebyla úplně pravda. Problémy se projevovaly v homopolymerních úsecích, kde „minus“ systém poskytoval spolehlivou informaci o začátku takového úseku a „plus“ systém o jeho konci. Přesné zjištění délky úseku ale nebylo spolehlivé ani při použití obou systémů společně, což bylo velkým nedostatkem této metody (Sanger a Coulson, 1975; Hutchison III, 2007).

3.2.3 Maxam-Gilbertova chemická metoda

V únoru 1977 publikovali Allan Maxam a Walter Gilbert sekvenační metodu založenou na chemickém štěpení DNA (Maxam a Gilbert, 1977). Sekvenovaná DNA je na jednom konci označena radioaktivním fosforem (^{32}P). Vzorek je pak rozdělen na čtyři části, každá z nich je podrobena jiné chemické reakci, napadající vždy určitou skupinu chemických bází. V místě poškozené báze je pak DNA chemicky rozštěpena. Protože ke štěpení dojde vždy jen na malé části z možných míst, je výsledkem směs radioaktivních fragmentů, jejichž délka odpovídá pozicím dané skupiny bází v sekvenované DNA. Fragmenty z každé ze čtyř reakcí jsou pak elektroforeticky rozděleny podle své délky na polyakrylamidovém gelu, zobrazeny autoradiograficky (fragmenty neobsahující radioaktivní fosfor nejsou na filmu zobrazeny a

neruší tak výsledek) a porovnáním výsledků ze čtyř reakcí lze přečíst sekvenci (Maxam a Gilbert, 1977).

Vzhledem ke své chemické podstatě nejsou čtyři používané chemické reakce specifické vždy pro jednu ze čtyř bází, ale jedná se o kombinace adenin+guanin (s vyšší citlivostí a tudíž výraznějšími pruhy pro guanin), adenin+guanin (s vyšší citlivostí pro adenin), cytosin+thymin a samotný thymin. Z těchto kombinací je však možné výslednou sekvenci bez problémů odečíst a dokonce poskytují určitou redundanci informace (Maxam a Gilbert, 1977).

Narozdíl od Sangerovy “plus a minus” metody z roku 1975, umožňovala Maxam-Gilbertova chemická metoda správné čtení homopolymerních úseků (produkuje správné pruhy i uvnitř těchto úseků). To byla velká výhoda a vedlo to po publikaci metody k jejímu širokému rozšíření (Hutchison III, 2007).

3.2.4 Sangerova dideoxymetoda

V roce 1977 publikoval Sanger novou metodu sekvenace, která vyřešila problémy s homopolymerními úseky, které měla jeho starší “plus a minus” metoda (Sanger et al, 1977). Principem metody je částečné nahrazení nukleotidů (dNTP) analogy neumožňujícími další prodlužování řetězce DNA polymerázou. V původní práci byly použity dva typy takových analogů, arabinonukleotidy (araNTP) a 2’,3’-dideoxyribonukleotidy (ddNTP). Arabinóza je stereoizomer ribózy, hydroxylová skupina na třetím uhlíku je v arabinóze v trans pozici vůči hydroxylové skupině na druhém uhlíku. To DNA polymeráze I z *Escherichia coli* překáží ve funkci a způsobuje ukončení řetězce (Sanger et al, 1977). Dodnes se jako analogy terminující řetězec využívají 2’,3’-dideoxyribonukleotidy, což jsou nukleotidy, kterým chybí hydroxylová skupina na třetím uhlíku deoxyribózy. Protože se k hydroxylové skupině na třetím uhlíku připojuje fosfodiesterovou vazbou další nukleotid, není další prodlužování řetězce zakončeného 2’,3’-dideoxyribonukleotidem vůbec možné (Sanger et al, 1977).

Sekvenace probíhá ve čtyřech samostatných reakcích. V každé ze čtyř reakčních směsí je použit kromě všech čtyřech normálních nukleotidů (dNTP) jeden ze čtyř dideoxynukleotidů (ddNTP). Dideoxynuklotidů je výrazně méně, v původní práci se uvádí poměr 1:100. Tak je v daném místě ukončena vždy jen malá část řetězců, většina pokračuje dále. Výsledkem každé

reakce je tak směs molekul o různé délce, vždy končící v místě nukleotidu, který odpovídá přidanému dideoxynuklotidu. Z reakce, do které byl přidán ddATP, tak získáme směs molekul o různé délce, ale vždy končící adeninem. Produkty všech čtyřech reakcí jsou, vedle sebe na jednom gelu, elektroforeticky rozděleny podle délky a zobrazeny autoradiograficky. Z gelu lze přímo odečíst sekvenci, s homopolymerními úseky nejsou problémy (Hutchison III, 2007).

Pro svou efektivitu a méně operací s toxickými a radioaktivními látkami (van Dijk et al., 2014) se Sangerova dideoxy metoda stala hlavní sekvenační technologií první generace a postupně byla zdokonalována (Liu et al., 2012). Původní značení radiokativním fosforem (^{32}P) bylo nahrazeno radioaktivní sírou (^{35}S), která díky nižší energii vyzařovaných částic produkuje ostřejší pruhy na autoradiogramu (Hutchison III, 2007). Ale metoda byla stále velmi pracná a vyžadovala práci s radioaktivními materiály (Liu et al., 2012).

3.2.4.1 První automatizované sekvenátory

První publikace o automatizované sekvenaci DNA vyšla v roce 1986. Jednalo se o modifikovanou dideoxymetodu bez použití radioaktivního značení. V každé ze čtyř reakcí byl použit primer označený jiným fluorescenčním barvivem. Výsledky reakcí byly smíchány a podrobeny kapilární elektroforéze v trubičce s polyakrylamidovým gelem. Detektor registroval barevné fluorescenční signály, které kolem něj procházely v pořadí odpovídajícím sekvenci DNA a data se nahrávala přímo do počítače (Hutchison III, 2007).

Automatické dideoxy sekvenátory umožnily v roce 1990 spuštění Projektu lidského genomu (Human Genome Project) financovaného vládou Spojených států amerických, jehož hlavním cílem bylo získat kompletní sekvenci lidského genomu. Paralelně se o totéž snažila soukromá společnost Celera, soutěž mezi oběma týmy sekvenování významně urychlila. První drafty sekvence od obou týmů byly publikovány počátkem roku 2001 v časopisech Science a Nature (Hutchison III, 2007).

3.2.5 Metody druhé generace

Před deseti lety se začaly objevovat sekvenační metody, které ohrozily výsadní postavení dideoxy metody. Jejich společnou vlastností je masivní paralelismus (odtud termín masivně paralelní sekvenování), tzn. že v jednom experimentu se najednou sekvenuje o mnoho více různých molekul, než bylo typické pro moderní kapilární sekvenátory využivající dideoxy

metodu (96 kapilár). Jednotlivé sekvenace pomocí těchto metod měly hlavně zpočátku nižší přesnost a umožňovaly číst kratší úseky DNA než Sangerova dideoxy metoda. Vzhledem k vysokému paralelismu (každý úsek je osekvenován vícekrát), však mohou být výsledná data velmi přesná (Hutchison III, 2007).

3.2.5.1 Pyrosekvenování (454/Roche)

První komerčně dostupná masivně paralelní sekvenovací technologie byla na trh uvedena firmou 454 Life Sciences (později zakoupená firmou Roche) v roce 2005. Metoda je založená na principu pyrosekvenování spočívajícím v bioluminiscenční detekci připojování nukleotidů během DNA syntézy. Uvolňovaný pyrofosfát je sekvencí enzymatických reakcí využíván k tvorbě světla, jehož intenzita je měřena (Metzker, 2010).

Nejprve je DNA rozbita na fragmenty vhodné délky, ke kterým jsou připojeny krátké sekvence (linkery), které umožní připojení každého jednotlivého fragmentu na vlastní mikrokuličku (bead) pokrytou primery. Jednotlivé mikrokuličky jsou pak izolovány v malých kapičkách PCR reakční směsi rozptýlených v oleji (emulze). Kapičky pak fungují jako samostatné PCR mikroreaktory (2×10^6/ml), čímž je umožněna paralelní ale oddělená amplifikace velkého množství fragmentů DNA. Na mikrokuličce je vytvořeno 10^7 kopií DNA fragmentu (Margulies et al., 2005). Přečištěné mikrokuličky s amplifikovanou DNA jsou pak po jedné umístěny do milionů pikolitrových jamek na speciální destičce s optickými vlákny (Hutchison III, 2007). Ta je umístěna do průtokové komůrky, která zajišťuje přívod potřebných reagencií (Margulies et al., 2005).

Vlastní sekvenační reakce probíhá paralelně v jednotlivých jamkách. DNA polymeráza I zbavená exonukleázové aktivity (Klenowovův fragment) syntetizuje s využitím primeru komplementární vlákno podle sekvenované DNA. Při připojení každého nukleotidu (dNTP) je uvolněn pyrofosfát (PP_i). Ten je enzymem ATP sulfurylázou využit k syntéze ATP. Vzniklý ATP je systémem luciferáza-luciferin (v přírodě se vyskytujícím například u světlušek) využit k produkci světla, které je elektronicky detekováno. Cyklus je ukončen apyrázou, která rozloží nevyužité nukleotidy a ATP (Ahmadian et al., 2006). Protože se v každém cyklu do reakční směsi přidá pouze jeden ze čtyř nukleotidů a protože pro homopolymerní úseky do délky šesti nukleotidů je světelný signál přímo úměrný počtu připojených nukleotidů (Metzker, 2010), je možné ze zaznamenaného pyrogramu odečíst sekvenci syntetizovaného

vlákna. Nicméně chyby (delece a inzerce) způsobené špatným vyhodnocením délky homopolymerních úseků jsou hlavním zdrojem chyb při tomto způsobu sekvenování (Hutchison III, 2007).

V roce 2005 byla délka sekvenovaného úseku 100-150 bází, v roce 2008 až 700 bází. Nevýhodou metody je vysoká cena reagencií (Liu et al., 2012). V roce 2013 firma Roche oznámila, že v roce 2016 ukončí podporu pro sekvenátory 454 (GenomeWeb, 2013).

3.2.5.2 Illumina (Solexa)

Tato metoda sekvenování byla uvedena na trh v roce 2006 firmou Solexa, která byla o rok později odkoupená firmou Illumina (Ansorge, 2009). Illumina tuto technologii nazývá prostě Sequencing By Synthesis (Illumina, 2015).

Během přípravy knihovny jsou k fragmentům DNA určeným k sekvenaci na obou stranách připojeny dva různé oligonukleotidy, tzv. adaptéry. Povrch průtokové komůrky je hustě pokryt oligonukleotidy komplementárními k oběma typům adaptérů (Shendure et al., 2011). Tyto oligonuklotidy slouží jako primery pro tzv. můstkovou PCR (během amplifikace vytváří vlákna DNA můstkové struktury mezi oligonukleotidy navázanými na stěnu komůrky (Metzker, 2010)). Výsledkem je, že produkty amplifikace zůstávají blízko sebe a vytvoří shluk obsahující zhruba 1000 kopií (Shendure a Ji 2008). Těchto shluků, které lze paralelně sekvenovat, jsou desítky milonů (Ansorge, 2009).

Reakční směs pro sekvenaci obsahuje kromě sekvenačního primeru a polymerázy i čtyři typy nukleotidů, z nichž každý je označen jiným fluorescenčním barvivem a navíc obsahuje terminační skupinu na svém 3’ uhlíku. Přítomnost terminační skupiny zaručí, že je do každého řetěze inkorporován pouze jeden nukleotid. Následně je elektronicky zaznamenán obraz, barevná fluorescence v místě každého z milionů sekvenovaných řetězců prozradí identititu posledního připojeného nukleotidu. Pak je odstraněna flurescenční značka i terminační skupina a následuje další cyklus syntézy (Ansorge 2009). V Ansorgově článku z roku 2009 se uvádí maximální sekvenovaná delká 35 nukleotidů, u současných přístrojů Illumina je to 300 bází (van Dijk et al., 2014).

3.2.5.3 Sekvenování ligací (ABI SOLiD)

V roce 2007 uvedla firma Applied Biosystems svůj sekvenační systém ABI SOLiD (Sequencing by Oligonucleotide Ligation and Detection), který je založen nikoliv na syntéze DNA polymerázou, ale na využití ligázy (enzymu katalyzujícího spojování řetězců DNA) (Ansorge, 2009).

Příprava knihovny je podobná jako v systému 454/Roche. Jednotlivé fragmenty jsou izolovány a amplifikovány na mikrokuličkách pomocí emulzní PCR. Mikrokuličky jsou ale menší, než v systému 454/Roche, asi 1 mikrometr v průměru. Mikrokuličky nejsou umístěny do mikrojamek, ale náhodně na povrch průtokové komůrky. Dosažená hustota je vyšší než v případě sytému 454/Roche nebo Illumina (Shendure et al., 2011).

Při vlastním sekvenování je použit univerzální primer, směs krátkých fluorescenčních nukleotidů a ligáza. Oligonuklotid nejlépe odpovídající sekvenci je ligázou připojen k primeru. Nepřipojené oligonukleotidy jsou odstraněny. Barva fluorescenční značky informuje nikoliv o celé sekvenci připojeného oligonukleotidu, ale pouze o jeho prvních dvou nukleotidech. Kamera sejme obraz. Fluorescenční značka je odstraněna a celý cyklus se opakuje (Shendure et al., 2011).

Oligonukleotidy používané k sekvenaci jsou dlouhé 8 nukleotidů, ale tři nukleotidy jsou v každém cyklu odstraněny spolu s fluorescenční značkou, DNA je tedy v každém cyklu prodloužena o pět nukleotidů. Po několika cyklech je vzniklé vlákno odstraněno a vše se opakuje s novým primerem, posunutým o 1 nukleotid, celkem pětkrát (Kircher a Kelso, 2010). Protože jsou použity jen čtyři různé barevné značky a možných dinukleotidů je 16, může každá barva představovat čtyři různé dinukleotidy. Protože je ale každá báze součástí dvou barevně označených dinukleotidů (každý při ligaci s jiným primerem), lze báze správně identifikovat. Toto tzv. dvoubázové kódování (2 base encoding) je navrženo matematicky tak, aby, je-li k dispozici referenční sekvence, umožňovalo detekci sekvenačních chyb (Kircher et Kelso, 2010; Breu, 2010).

3.2.5.4 Iontové polovodičové sekvenování (Ion Torrent)

Tato sekvenační technologie, uvedená na trh v roce 2010 firmou Ion Torrent (Rusk, 2011), je založena na detekci protonů (vodíkových iontů, H^+) uvolňovaných při polymeraci DNA z 3' hydroxylové skupiny na rostoucím řetězci (Merriman et al., 2012).

Základní princip polovodičového iontového sekvenování se podobá pyrosekvenování v tom smyslu, že se střídají cykly se čtyřmi typy nemodifikovaných nukleotidů a detekuje se, ve kterých cyklech byl nukleotid (nebo více nukleotidů) připojen. Detekce ovšem není založena na bioluminiscenci, ale měří se vodíkové ionty (H^+) uvolňované při polymerační reakci (tedy vlastně změny pH). Je-li v daném cyklu připojen nukleotid, dojde ke změně pH. Je-li připojeno více stejných nukleotidů, změna pH je větší (Shendure et al., 2011).

V roce 2010 byla délka sekvenovaného úseku typicky 100 bází, v roce 2012 400 bází. Nejčastější chyby jsou podobně jako u pyrosekvenace způsobeny špatným vyhodnocením délky homopolymerních úseků (Merriman et al., 2012)

3.2.6 Sekvenování jednotlivých molekul

Nejnovější sekvenovací metody (někdy nazývané metody třetí generace, i když toto označení se používá nejednotně) umožňují sekvenování jednotlivých molekul DNA (Single Molecule Sequencing, SMS). To znamená, že sekvenování nepředchází amplifikace DNA pomocí PCR. Nesekvenuje se shluk identických molekul, ale molekula jediná. Nemůže tedy docházet k ovlivnění výsledku kvůli PCR amlifikaci (PCR bias) a nevznikají problémy se synchronizací sekvenace (dephasing), což problémy metod druhé generace (Schadt et al., 2010).

3.2.6.1 Helicos Biosciences

První komerčně dostupná metoda využívající sekvenování jednotlivých molekul byla uvedena firmou Helicos Biosciences v roce 2007 (Ansorge, 2009). Fragmenty DNA jsou denaturovány a k jejich 3' konci je připojeno několik adeninů (poly(dA) tail) a fluorescenční značka. Tím je vytvořena knihovna pro sekvenaci. Fragmenty jsou hybridizovány s oligo(dT) molekulami, které jsou v miliardových počtech kovalentně připojeny ke skleněnému povrchu v průtokové komůrce. Pak jsou do komůrky střídavě přiváděny čtyři typy fluorescenčně

značených nukleotidů, řetězec DNA je prodlužován polymerázou a v každém cyklu je vždy zaznamenán obraz a následně odstraněna fluorescenční značka (Shendure et al., 2011) i skupina inhibující připojování dalších nukleotidů (Metzker, 2010). Protože je proces v každém kroku přerušován podobně jako u metod druhé generace, čas pro osekvenování každého nukleotidu je poměrně velký. Délka sekvenovaného úseku je omezená na asi 32 nukleotoidů a chyby v datech získaných z jednoho řetězce mohou být vyšší než 5%, což je ale kompenzováno vysokou paralelizací, takže výsledná přesnost sekvence po zpracování (konsenzus) přesahuje 99% (Schadt et al., 2010).

V roce 2012 požádala Helicos Biosciences o ochranu před věřiteli (GenomeWeb, 2012).

3.2.6.2 Pacific Biosciences SMRT

Technologie SMRT (Single Molecule Real-Time Sequencing, tj. sekvenování jedné molekuly v reálném čase) je první metodou sekvenace, při které je přímo pozorována činnost DNA polymerázy během připojování jednotlivých nukleotidů do vznikajícího řetězce. Protože DNA polymeráza je velmi malá (řádově 10 nm), musí být pro zachování dostatečného odstupu signálu od šumu objem, ve kterém pozorování probíhá, také velmi malý. K tomu se využívají tzv. zero-mode waveguides (ZMW), což jsou v podstatě otvory o průměru desítek nanometrů v tenkém (100 nm) kovovém filmu naneseném na skleněném podkladu (Schadt et al., 2010).

Princip je ZMW podobný jako princip ochranné mřížky v mikrovlnné troubě, kde mikrovlny, dlouhé řádově desítky centimetrů, neproniknou řádově milimetrovými otvory v mřížce, kterými však prochází viditelné světlo umožnující pohled do nitra trouby. Průměr ZMW je menší než vlnová délka viditelného laserového světla používaného k pozorování (cca 600 nm). Laser tak osvítí jen spodních 30 nm ZMW, dál se světlo nedostane (Schadt et al., 2010).

Ke skleněnému dnu ZMW je pomocí biotinu a streptavidinu připojena DNA polymeráza. Fluorescenčně označené (každý typ nukleotidu je značen jinou barvou) nukleotidy difundují skrz ZMW k polymeráze, která je může využívat k syntéze DNA. Nukleotidy, které nejsou polymerázou využity rychle difundují pryč (difúze trvá řádově mikrosekundy), zatímco připojování nukleotidu trvá řádově milisekundy (Schadt et al., 2010). Protože je barvivo

připojeno k fosfátové skupině, je během polymerace odštěpeno a difunduje pryč (Shendure et al., 2011). Tak je možno zaznamenávat silný signál připojovaných nukleotidů.

Protože se při této metodě sleduje kinetika polymerázy a protože je kinetika polymerázy ovlivňována modifikacemi DNA předlohy (methylace apod.), je možné při sekvenování pomocí technologie SMRT zjišťovat nejen sekvenci DNA, ale i její epigenetické modifikace (Flusberg et al., 2010).

První komerční verze SMRT sekvenátoru používala pole 75000 ZMW (Schadt et al., 2010), dnes 150000 ZMW (Pacific Biosciences, 2014), v každém ZMW může být DNA polymeráza s jiným DNA vláknem, které se tak sekvenují paralelně (Schadt et al., 2010).

3.2.6.3 Další metody sekvenování jednotlivých molekul

V různém stupni vývoje je řada dalších sekvenačních metod založených na sekvenování jednotlivých molekul DNA. Jednou skupinou jsou sekvenační metody založené na přímém pozorování DNA molekul pomocí neoptické mikroskopie - transmisní elektronové mikroskopie nebo řádkovací tunelové mikroskopie (Schadt et al., 2010).

Další skupina metod využívá pro sekvenaci nanopóry, a to buď vytvořené z proteinů upravených genovým inženýrstvím nebo zcela syntetické. Velkou výhodou je přímé sekvenování DNA nebo RNA bez složité přípravy knihovny nebo sekvenačních reagencií (van Dijk et al., 2014) a bez nutnosti značení jednotlivých nukleotidů (Thompson et Milos, 2011). Firma Oxford Nanopore vytvořila systém využívající póry tvořené alfa hemolysinem (toxin bakterie *Staphylococcus aureus*) v syntetické lipidové membráně. Na membráně je vytvořeno napětí. K vnější části póru je připojena exonukleáza, která postupně odštěpuje jednotlivé nukleotidy, které procházejí pórem a při průchodu charakteristickým způsobem narušují proud iontů skrze pór. Změny proudu jsou elektronicky měřeny (Schadt et al., 2010).

Novější metoda od Oxford Nanopore využívá průchod celé intaktní molekuly DNA pórem. V roce 2014 uvedli MinION, paralelní (několik set pórů) sekvenátor velikosti mobilního telefonu, založený na této metodě. Umožňuje sekvenovat tisíce bazí dlouhé úseky a celý genom *Escherichia coli* lze přečíst během jednoho běhu přístroje (Bayley, 2015).

3.2.7 Srovnání důležitých sekvenačních metod

V současnosti (k roku 2014) je nejdůležitější sekvenační metodou Illumina, která nabízí maximální průchodnost a nejnižší cenu za osekvenovanou bázi. Vlastnosti, výhody a nevýhody hlavních sekvenačních metod nové generace shrnují následující tabulky:

Platforma	**Max. délka čtení (nukleotidy)**	**Max. průchodnost (Gb za běh)**	**Délka běhu (hodiny)**
454/Roche	1000	0,7	10
Illumina	300	1800	27
SOLiD	75	320	336
Ion Torrent	400	10	3
PacBio SMRT	20000	0,5	3

Tabulka 1: Srovnání metod nové generace s významným zastoupením na trhu. Délka běhu je měřena při "typickém sekvenování bakteriálního genomu". Uvedené parametry nejsou vždy od stejného přístroje (modelu). Zobrazeny nejlepší aktuální hodnoty z roku 2014 (van Dijk et al., 2014).

Platforma	**Výhody**	**Nevýhody**
454/Roche	dlouhá délka čtení (1 kb) relativně rychlý běh	relativně malá průchodnost vysoká cena reagencií vysoká chybovost v homopol. úsecích ukončení podpory v roce 2016
Illumina	leader na trhu nejvyšší průchodnost nejnižší cena za bázi	technicky náročné, koncentrace vzorku je kritická, překrývající se klastry snižují kvalitu čtené sekvence
SOLiD	druhá nejvyšší průchodnost nízká chybovost	nejkratší délka čtení (75 nt) pomalý běh méně vhodné pro de novo sekvenování
Ion Torrent	nevyužívá optiku a fluorescenci velmi rychlý běh	vysoká chybovost v homopol. úsecích
PacBio SMRT	velmi dlouhá délka čtení (20 kb) velmi rychlý běh	vysoká cena vysoká chybovost nejmenší průchodnost

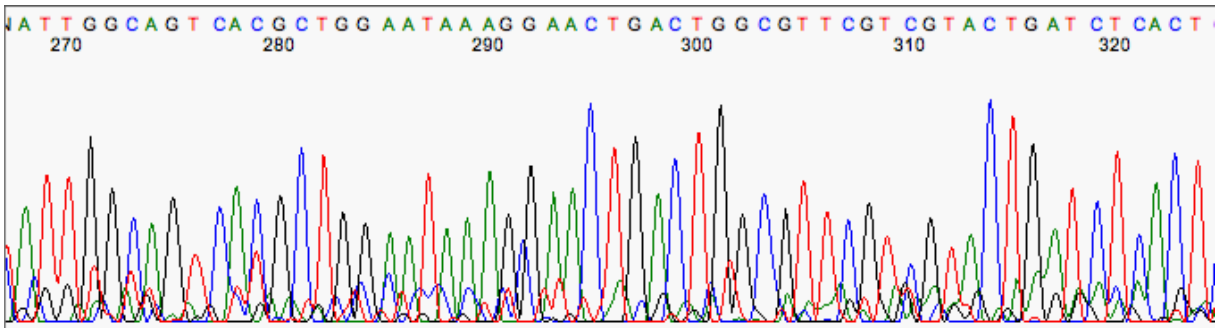
Tabulka 2: Výhody a nevýhody sekvenovacích metod nové generace s významným zastoupením na trhu (van Dijk et al., 2014).

3.3 Zpracování sekvenačních dat

3.3.1 Určování bází (base-calling) a skládání sekvencí

Hrubá data vycházející ze sekvenátoru nemají podobu konkrétní jednoznačné sekvence. V případě Sangerových sekvenátorů, kde jsou jednotlivé báze značené čtyřmi různými fluorescenčními barvičkami, má výstup, trace, podobu chromatogramu, tj. grafu se čtyřmi barevnými křivkami, z nichž každá představuje průběh signálu pro jednu z bází (respektive odpovídající barvičku) v čase. V ideálním případě by se mělo jednat o jasně oddělené nepřekrývající se vrcholy jasně určující sekvenci. Z řady důvodů reálné chromatogramy nevypadají takto ideálně a obsahují šum, zejména na začátku a konci chromatogramu, ale v menší míře i v celém jeho průběhu. Base-calling (určování bází) je proces překladu hrubých dat ze senzoru do sekvence bází (Ewing et al., 1998).

Příklad méně kvalitního chromatogramu obsahujícího více šumu (převzato z University of Michigan, 2015):



Známý program, který se používá pro určování bází z chromatogramů, se nazývá Phred (Ewing et al., 1998). Tzv. Phred quality score, přiřazený jednotlivým bázím, udává pravděpodobnost chybného určení báze (jedná se o logaritmickou míru) (Ledergerber et Dessimoz, 2011).

Base-calling se používá i u modernějších metod sekvenace, jako je pyrosekvenace (454), Illumina a dalších. U jednotlivých metod se liší technické provedení, typy senzorů a tudíž i typy chyb, které se v datech vyskytují. Přesnost sekvenování lze zlepšit zvýšením pokrytí

(coverage), tedy resekvencováním stejných úseků DNA několikrát. Ale přesnější base-calling software dokáží zvýšit přesnost bez zvyšování pokrytí a tím mohou snížit cenu sekvenování (Ledergerber et Dessimoz, 2011).

Pokud je délka čtených úseků ze sekvenátoru kratší, než sekvence, která nás zajímá, je nutno sekvenci během zpracování v počítači sestavit z překrývajících se úseků. Tento krok se nazývá skládání sekvence (sequence assembly). Je zvláště obtížný při de novo sekvenování celého genomu technologií, která poskytuje krátké čtené úseky. Důležitá je hloubka pokrytí (depth of coverage), tedy kolikrát je každá báze (v rámci různých fragmentů) osekvenována. Například sekvenování různých genomů pomocí technologie Illumina s délkou čtených úseků 35-100 nukleotidů využívalo 50-100násobné pokrytí, zatímco při klasickém Sangerově sekvenování s délkou čtených úseků vyžadovalo pouze 7-10násobné pokrytí. Nicméně skládání z krátkých úseků představuje velmi obtížný problém (Salzberg et al., 2011).

Software, který řeší problém skládání sekvencí, se nazývá assembler. Mezi významné, opensource assemblery patří například ABySS, ALLPATHS-LG, Bambus2, CABOG, MSR-CA, SGA, SOAPdenovo nebo Velvet (Salzberg et al., 2011).

3.3.2 Databáze a vyhledávání sekvencí

Existují tři hlavní světové databáze DNA sekvencí. Je to GenBank, provozovaná Národním centrem pro biotechnologické informace v USA a dostupná na na adrese http://www.ncbi.nlm.nih.gov/genbank/, EMBL ve Velké Británii, která je součástí European Nucleotide Archive (http://www.ebi.ac.uk/ena) a DDBJ v Japonsku (http://www.ddbj.nig.ac.jp/). Sekvence jsou do databáze GenBank vkládány skrze webové rozhraní nebo specializovaným programy (Sequin) (Benson et al., 2013). Databáze jsou navzájem propojené, takže zanesení sekvence do jedné z nich se odrazí ve všech databázích (Macholán, 2014). Datová výměna s ENA a DDBJ probíhá denně (Benson et al., 2013). Krom toho existují specializované databáze, jako například databáze BOLD pro DNA barcoding (Ratnasingham et Hebert, 2007).

Množství dat v databázi GenBank se každých 10 měsíců zdvojnásobuje (Macholán, 2014). Dlouhou dobu vývoj rychlosti a kapacity počítačů (známý Mooreův zákon, říkající, že

množství tranzistorů na integrovaném obvodu se zdvojnásobuje zhruba každých 18 měsíců) předčil vývoj sekvenovacích technologií. Podobně se i kapacita počítačových disků dlouhou dobu zvyšovala rychleji. Rozvoj sekvenačních technologií nové generace ale tento trend změnil, a nyní se sekvenování vyvíjí rychleji (osekvenovaná báze za jednotku ceny) než kapacita disků (velikost paměti za jednotku ceny). S rozvojem metod druhé generace klesá cena sekvenování jedné báze o polovinu každých pět měsíců. To je výrazně rychlejší, než rozvoj výpočetní techniky a uložení a zpracování sekvenačních dat se tak stává významným technickým problémem (Stein, 2010).

Každá sekvence v databázi má své unikátní přístupové číslo (accession number), podle kterého lze jednoznačně identifikovat. Toto číslo je sdílené mezi GenBank, DDBJ i EMBL. Přístupové číslo se nemění ani při aktualizaci sekvence (vždy ukazuje na nejaktuálnější verzi). V případě aktualizac sekvence se mění její číslo verze a identifikátor GI (Benson et al., 2013). K sekvenci je dále připojena řada informací, jako jméno lokusu, druh organismu, zdroj sekvence a další (Macholán, 2014). Databáze ale obsahuje i sekvence patřící neznámým organismům, například sekvence ze vzorků prostředí (Benson et al., 2013).

Vlastní sekvence mohou být uloženy v různých formátech. Jednoduchým formátem využívaným i v GenBank je FASTA. Před každou sekvencí je řádek začínající znakem ">", za kterým může být identifikátor a popřípadně i popis sekvence, na dalších řádcích je pak vlastní nepřerušená (respektive přerušená pouze odřádkováním po typicky maximálně 80 znacích) sekvence. V jednom textovém souboru může být takových sekvencí více. Existuje ale řada dalších formátů specifických pro jednotlivé programy pracujícími se sekvencemi. Některé programy (např. PHYLIP a PAUP*) umožňují pracovat se sekvencemi v proloženém formátu. Proložený formát znamená, že jsou přímo pod sebou vždy odpovídající řádky všech sekvencí v souboru, což zvyšuje přehlednost, jedná-li se o podobné sekvence (Macholán, 2014).

Pro vyhledávání sekvencí v databázi podle jiné sekvence se používá algoritmus BLAST (existující v několika modifikacích). To lze využít například, chceme-li zjistit, k jakému organismu či genu náleží neznámá sekvence. Algoritmus nejprve hledanou sekvenci rozdělí na menší úseky a hledá shodu v databázi, nalezne-li podobnou sekvenci, hledání se od tohoto

úseku rozšiřuje v obou směrech. Výsledkem hledání je seznam nejpodobnějších sekvencí, seřazených podle míry shody (Macholán, 2014).

3.3.3 Alignment

Důležitou operací se sekvencemi, která má velký význam mimo jiné ve fylogenetice (a užívanou i při výše popsaném skládání a vyhledávání sekvencí) je alignment (seřazení). Je to základní krok i při konstrukci fylogenetického stromu. Máme-li dvě, nebo obecně více, příbuzných sekvencí, které potřebujeme porovnat, potřebujeme je vedle sebe položit tak, aby na stejných pozicích byly nukleotidy, které pocházejí od společného předka těchto sekvencí. Neshody mezi sekvencemi mohou být důsledkem buď substitučních mutací, nebo inzercí a delecí (Macholán, 2014). Všechny části všech genomů totiž vznikly z jednoho původního genomu pomocí různých mutací. (Felsenstein, 2014). Při alignmentu tedy hledáme takový způsob seřazení sekvencí, aby bylo možno rozdíly mezi nimi vysvětlit co nejmenším počtem substitucí, delecí či inzercí. Důsledkem delecí a inzercí jsou mezery v seřazených sekvencích. S využitím příliš velkého množství mezer ale může alignment ztrácet biologický smysl. Proto jsou mezery často vzhledem k substitucím penalizovány (gap penalty). Míra penalizace mezer by měla odrážet frekvenci inzercí a delecí (Macholán, 2014).

Asi nejrozšířenějšími programy pro alignment sekvencí je rodina programů Clustal (Felsenstein, 2004). Zejména Clustal Omega dosahuje vysokých rychlostí i při alignmentu velkého množství sekvencí (Macholán, 2014). Program je k dispozici zdarma a má i online webové rozhraní (http://www.ebi.ac.uk/Tools/msa/clustalo/).

3.4 DNA barcoding

Velká část biologického výzkumu vyžaduje identifikaci biologických druhů. Klasická identifikace založená na morfologických znacích má však řadu nevýhod. Aby bylo možné identifikovat libovolný z milionů biologických druhů, je potřeba velké množství specialistů zaměřených na jednotlivé taxonomické skupiny. Fenotypická plasticita a morfologicky kryptické druhy mohou vést ke špatné identifikaci. Používané určovací znaky jsou mnohdy použitelné jen na určitá vývojová stádia nebo konkrétní pohlaví. Navíc identifikace na základě morfologie vyžaduje značné zkušenosti, takže poměrně běžně dochází k nesprávné druhové identifikaci (Hebert et al., 2003).

Tyto problémy se v plné míře týkají i ichtyologie, zejména jedná-li se o různá vývojová stádia a nebo například zpracované ryby či jejich části (Ward et al., 2009). Proto se začalo pro identifikaci druhů využívat i molekulárních znaků, nejprve allozymové analýzy (elektroforetické rozlišení různých alelických forem proteinů, 1975), později i mitochondriální DNA. Identifikace pomocí DNA má oproti proteinům řadu výhod, DNA je méně náchylná k degradaci, je přítomná ve všech vývojových stádiích, pomocí PCR amplifikace lze analyzovat i velmi malé vzorky tkáně a k identifikaci lze využívat i synonymních mutací, které se v proteinech neprojeví (Ward et al., 2009). V roce 1991 byla použita mitochondriální DNA (gen cytochromu b) pro identifikaci čtyř různých druhů tuňáků (*Thunnus*) (Ward et al., 2009).

Nejprve byly k identifikaci druhů využívány různé části různých genů. V roce 2003 Hebert a spolupracovníci navrhli metodu, tzv. DNA barcoding, která využívá jeden úsek jednoho genu jako základ pro univerzální identifikační systém pro všechny živočichy (Ward et al., 2009). Název DNA barcoding je analogií k čárovému kódu (angl. barcode), který se používá k identifikaci zboží (Hebert et al., 2003). Základní předpokladem pro DNA barcoding je myšlenka, že DNA každého biologického druhu obsahuje krátkou sekvenci DNA, která je pro daný druh specifická, tedy funguje jako jakýsi otisk prstu. Tato sekvence by měla pocházet z části genomu, která se vyvíjí dostatečně rychle, aby bylo možné rozlišit blízce příbuzné druhy (sdílející společného předka), ale zároveň aby byly co nejmenší rozdíly mezi příslušníky stejného druhu (Roopnarine 2006).

3.4.1 Cytochrom c oxidáza

V případě živočichů byla jako vhodná univerzální sekvence pro DNA barcoding zvolena část sekvence mitochondriálního genu pro podjednotku I cytochrom c oxidázy (COI), enzymu, který je součástí dýchacího řetězce v mitochondriích. Část genu (zhruba 650 bází na 5' konci (Ward, 2009)) je amplifikována pomocí PCR (s využitím univerzálních primerů), amplifikovaná DNA je následně osekvenována a porovnána s databází. Výhoda mitochondriální DNA spočívá v nepřítomnosti intronů, minimální rekombinaci a haploidní dědičnosti, což zjednodušuje analýzu. Rychlost evoluce COI je navíc zrhuba třikrát rychlejší než mitochondriální ribozomální RNA (12S a 16S). (Hebert, 2003). Vnitroduhová variabilita

tohoto genu je nízká ve srovnání s variabilitou mezidruhovou, pro jednotlivé druhy tedy obvykle existuje buď specifická sekvence nebo shluk velmi podobných sekvencí (Ward, 2009). Tato standardizace byla důležitým krokem k vývoji metody a rozvoji databází (Taberlet et al., 2012).

Využití mitochondriálního genu (COI) pro DNA barcoding má ale i své nevýhody. Protože se mitochondriální DNA dědí v naprosté většině případů výhradně po matce, jsou případní mezidruhoví kříženci (u ryb i v přírodě známý jev) identifikování jako druh mateřský. V takovém případě je pro správnou identifikaci nutno použít i jadernou DNA. Tento problém je ale poměrně vzácný (Ward et al., 2009). Dalším problémem je, že pomocí COI nelze odlišit některé teprve nedávno vzniklé druhy. Příkladem jsou například cichlidy (Cichlidae) z velkých východoafrických jezer, kde došlo ke vzniku mnoha nových druhů teprve nedávno. V takových případech může být nutné použít rychleji se vyvíjející gen (Ward et al., 2009). Dalším problémem může být heteroplasmie (výskyt různých mitochondriálních haplotypů v jednom organismu) případně výskyt nukleární mitochondriální DNA (pseudogeny mitochondriálního původu v jádře) (Taylor et Harris, 2012).

3.4.2 Barcoding iniciativy a databáze

Barcode of Life Data System (BOLD, http://www.barcodinglife.org/) je web obsahující centrální databázi DNA barcoding sekvencí (veřejně dostupná data, i data dosud nepublikovaná, chráněná heslem), primerů a nástroje pro správu a analýzu dat (Ratnasingham et Hebert, 2007). Obsahuje například nástroj pro identifikaci živočišných druhů na základě 5' části sekvence COI. Nejlépe odpovídající záznam je vyhledáván pomocí kombinace algoritmu BLAST (Basic Local Alignment Search Tool) a skrytého markovského modelu (Hidden Markov Model, HMM) (Bold Systems, 2014). Databáze obsahuje sekvence a identifikační nástroje i pro další skupiny organismů (rostliny a houby, pro které se používají jiné geny) (Bold Systems, 2014).

Iniciativa FISH-BOL (The Fish Barcode of Life Initiative, http://www.fishbol.org), založená roku 2005 a fungující v návaznosti na BOLD, si klade za cíl shromáždit záznamy pro DNA barcoding všech druhů ryb (od každého druhu pokud možno více exemplářů). Správná sekvence pro DNA barcoding musí obsahovat nejméně 500 párů bází z 5' oblasti

COI. Měla by být doplněna informacemi o dokladovém exempláři (voucher specimen - který by měl být uložen ve veřejně dostupné sbírce), zeměpisnými souřadnicemi lokality sběru, datu sběru, a identitou sběratele a člověka, který exemplář určil. Pokud není možné uchovat dokladové exempláře (velké ryby, ohrožené druhy), měl by být záznam doplněn alespoň fotografiemi. Důležitou infromací jsou také použité PCR primery a nezpracovaná sekvenační data (trace archive dostupný na adrese http://www.ncbi.nlm.nih.gov/Traces/) (Ward, 2009).

3.4.3 Identifikace rybích produktů

DNA barcoding umožňuje identifikaci všech vývojových stádií i fragmentů ryb, včetně ryb zpracovaných nebo tepelně upravených (Ward, 2009). To lze využít pro identifikaci konzumních ryb a rybích produktů na trhu. Například studie z irského Dublinu pomocí DNA barcodingu (COI) zjistila 25% druhově špatně označených rybích produktů. Studie se zaměřila na tresku obecnou (*Gadus morhua*) a tresku skvrnitou (*Melanogrammus aeglefinus*). 23.7% (37 ze 156) produktů bylo identifikováno jako druhy z jiných rodů než *Gadus* a *Melanogrammus*, zejména jako treska tmavá (*Pollachius virens*) a treska polak (*Pollachius pollachius*). Nejčastěji byly špatně označeny uzené produkty (82.4%, 28 z 34) (Miller et Mariani, 2010).

Podobných studií existuje celá řada, např. Wong a Hanner (2008), Filonzi et al. (2010), Carvalho et al. (2011) nebo Galal-Khallaf et al. (2014). Záměna ryb může být úmyslným podvodem, kdy jsou méně hodnotné druhy prodávané jako druhy hodnotnější, a někdy i zdravotně nebezpečná. Příkladem takového zdravotního rizika je lates nilský (*Lates niloticus*), který je často zneužíván jako náhražka jiných druhů ryb. Lates nilský pocházející z afrických řek bývá totiž kontaminován methylrtutí, kumulativním toxinem, který zvyšuje riziko infarktu myokardu a neurologických postižení (Filonzi et al., 2010). Nebezpečná může být rovněž záměna toxického čtverzubce za ďasa (Galimberti et al., 2013). V Evropě a Severní Americe dosahují podvody v případě mořských potravin 15-43%, zvlášť hodně případů (75%) se týká chňapala červeného (*Lutjanus campechanus*) (Galimberti et al., 2013).

3.4.4 Barcoding a skladba kořisti

DNA barcoding lze využít i pro studium potravy, kterou se ryby živí. Dunn et al. (2010) například využili DNA barcoding (COI) společně s vizuální identifikací pro analýzu obsahu žaludků několika druhů žraloků ulovených vlečnou sítí východně od Nového Zélandu. Použití

DNA barcodingu umožňuje identifikaci kořisti i u žraloků, které vykusují kusy z větší kořisti (např. světloun Bonnaterrův, *Dalatias licha*), a vizuální identifikace je proto obtížná. Jiným příkladem je studie skladby kořisti u okounka pstruhového (*Micropterus salmoides*) v závislosti na jeho velikosti, kterou publikoval Jo et al. (2014). S využitím DNA barcodingu (COI) identifikovali 26 typů kořisti ze čtyř kmenů, z toho 15 na úroveň druhu. Perutýn ohnivý (*Pterois volitans*) je nebezpečně se šířící dravou invazní rybou na Mezoamerickém korálovém útesu, což je druhý největší korálový útes na světě. Valdez-Moreno et al. (2012) publikovali studii skladby kořisti v žaludcích pomocí DNA barcodingu (COI). V žaludcích bylo identifikováno 34 rybích druhů, zejména z čeledí Gobiidae (hlaváčovití) a Apogonidae (parmovcovití). Byl potvrzen kanibalismus. Zjištěno bylo i 20 různých korýšů, z toho 12 desetinožců, ale jen jednoho bylo možné s využitím databází BOLD a GenBank určit do druhu. Riemann et al. (2010), s využitím genu pro 18S ribozomální RNA (jaderný gen), analyzoval potravu malých (4,5-14,5 mm) larev úhoře říčního (*Anguilla anguilla*) v Sargasovém moři. Bylo zjištěno, že důležitou potravou úhořích larev v této velikosti je rosolovitý zooplankton (polypovci - Hydrozoa, salpy - Thaliacea a žebernatky - Ctenophora).

3.4.5 Barcoding a importy ryb

DNA barcoding lze využívat i pro kontrolu druhů v mezinárodním obchodě. Akvarijní ryby jsou z hlediska importu nejdůležitější skupinou obratlovců. Regulace obchodu s akvarijními rybami je proto důležitá jak z hlediska ochrany druhů před nadměrným lovem ve své domovině (Steinke et al., 2009) tak i z hlediska šíření nebezpečných invazivních druhů a přenašečů patogenů. Různé státy řeší tento problém různými způsoby, buď seznamy zakázaných druhů (blacklist) nebo naopak povolených (whitelist). Vymáhání těchto pravidel je však obtížné kvůli nesnadné identifikaci druhů a také kvůli možnému přimíšení dalších druhů do zásilky (Collins et al., 2013). Identifikaci druhů může usnadnit DNA barcoding (Collins et al., 2012).

Problémem je ale odběr vzorku DNA. V případě cenných ryb není možné obětovat celého jedince a i malý odběr tkáně například z ploutve ryby může znamenat riziko infekce. Navíc je problém samotný výběr jedinců k odběru vzorků. Pozornosti totiž mohou uniknout odlišné, ale podobné, druhy přimíšené do zásilky. Možné řešení obou problémů spočívá ve využití extracelulární environmentální DNA (eDNA), tedy DNA přítomné ve vzorku odebraném

z prostředí, například vody. Tato technika dokonce umožňuje detekci druhů, které již v zásilce nejsou přítomny, ale byly v kontaktu s importovanými rybami dříve, například u obchodníka. Pokud se jedná o rizikový druh, mohou se na základě takovéto informace učinit potřebná karanténní opatření. Nicméně protože je DNA ve vodě nestabilní a podléhá degradaci, je při využití eDNA nutno pracovat s kratšími markerovými sekvencemi než při klasickém barcodingu (Collins et al., 2013).

3.4.6 Metabarcoding

Metody pro automatickou identifikaci více druhů (či vyšších taxonů) z jednoho vzorku, například vzorku z prostředí s degradovanou eDNA (voda, půda, výkaly a podobně), se označují jako DNA metabarcoding. Lze využívat čerstvé i velmi staré vzorky. Degradovaná DNA znemožňuje spolehlivé využití klasických sekvencí využívaných pro barcoding a vyžaduje kratší úseky DNA specificky navržené pro danou studii. Je potřeba krátká variabilní sekvence ohraničená dvěma vysoce konzervativními úseky okolo 20 párů bází pro připojení primerů. Ve srovnání s dlouhou sekvencí při klasickém DNA barcodingu je taxonomické rozlišení nižší. Proto je pro různé skupiny organismů (bakterie, členovci, obratlovci) vyžadována samostatná PCR, což znesnadňuje porovnání zastoupení mezi jednotlivými skupinami (Taberlet et al., 2012). Problémem je i tvorba vhodných databází, která je velmi nákladná. Možným řešením je využití kratších sekvencí v rámci sekvencí používaných pro klasický barcoding, což je ale obtížné. Zdrojem chyb je i amplifikace pomocí PCR. Ta by mohla být nahrazena vychytáváním vhodných fragmentů DNA pomocí k tomu navržených oligonuklotidů, případně přímé sekvenování eDNA extraktu bez amplifikace hledaných úseků pomocí moderních vysoce paralelních sekvenačních metod (např. Illumina) (Taberlet et al., 2012).

3.4.7 Omezení barcodingu

Důležité je, že barcoding je metoda pro identifikaci již známých druhů. Sekvence jednoho genu, navíc pouze částečná, není sama o sobě dostatečná pro popis nových druhů nebo tvorbu fylogenetických stromů (Taylor et Harris, 2012).

3.5 Molekulární fylogenetika

Fylogenetika je obor studující fylogenezi, tedy vznik a vývoj jednotlivých vývojových linií organismů, historii evoluce života na Zemi. Snaží se rekonstruovat průběh kladogeneze, tedy pořadí a způsob odvětvování (vzácněji i splývání - symbiogenezí nebo mezidruhovou hybridizací) jednotlivých vývojových linií, a opírá se o studium anageneze, tedy vývoje jednotlivých vlastností organismů v rámci vývojových linií (Flegr, 2009). V rámci procesu kladogeneze dochází k tzv. speciačním událostem, kdy z jednoho druhu mateřského vzniknou dva druhy dceřinné. Ke speciacím může docházet velmi rychle po sobě, ale je nepravděpodobné, že se více druhů odštěpilo skutečně ve stejný okamžik. Proto je v rámci rekonstrukce kladogeneze snaha proces vyjádřit jako sekvenci binárních větvení a pouze pokud to dostupná data neumožnují, zahrnují se multifurkace (mnohonásobná větvení) (Flegr 2009). Předpoklad binárních větvení je ale kontroverzní (Brinkman a Leipe, 2001).

3.5.1 Fylogenetické stromy

Fylogenetická analýza je vyvozování nebo odhadování těchto fylogenetických vztahů (Brinkman a Leipe, 2001). Výsledkem fylogenetické analýzy je fylogenetický strom, skládající se z větví (angl. branches, edges) a uzlů (angl. nodes, vertices). Uzly jsou terminální (externí) a vnitřní (interní), větve končící terminálním uzlem jsou periferní (koncové, terminální), větve spojující dva vnitřní uzly jsou vnitřní (Macholán 2004; Flegr, 2009). Koncové větve stromu představují druhy, které jsme zahrnuli do analýzy, vnitřní větve, vzhledem k omezené době trvání druhů, představují zpravidla již vymřelé předky. Pokud jsme do analýzy zahrnuli jen žijící druhy, jedná se o předky hypotetické, jejichž vlastnosti jsme odvodili od druhů dnešních. Jsou-li do analýzy zahrnuty i druhy vyhynulé, známé z paleontologického záznamu, lze je umístit i na vnitřní větve. Délka větví fylogenetického stromu vyjadřuje dobu trvání jednotlivých druhů či množství evolučních změn. V případě, že délka větví grafu nevyjadřuje žádný biologický význam, jedná se pouze o schéma kladogeneze (Flegr, 2009). Ve stromu podle jeho matematické definice (je necyklický) dochází pouze ke štěpení, nikoliv spojování větví. Je-li potřeba vyjádřit i splývání linií, je

potřeba použít síť (Macholán, 2014). Strom, který obsahuje pouze bifurkace, se nazývá binární neboli plně vyřešený (angl. fully resolved) (Macholán, 2004).

Fylogenetický strom může být buď s kořenem (angl. rooted tree), tzn. že je identifikován nejstarší společný předek ostatních taxonů, nebo bez kořene (angl. unrooted tree), kde nejstarší společný předek identifikován není (Macholán, 2004). Strom, jehož terminální uzly jsou pojmenovány, ze nazývá označený (angl. labeled). Kořen fylogenetického stromu lze stanovit, přidáme-li do analýzy tzv. vnější skupinu (angl. outgroup), která by neměla být příliš fylogeneticky vzdálená od studovaných taxonů (tzv. vnitřní skupina, ingroup), ale nesmí být její součástí. V ideálním případě by to měla být skupina sesterská (Macholán 2014).

3.5.2 Molekulární znaky využívané pro fylogenetickou analýzu

V dobách před rozvojem molekulární biologie se pro fylogenetickou analýzu používaly klasické, především morfologické znaky. V poslední době se stále více využívá znaků molekulárních, a to i v ichtyologii (Carvalho et Pitcher, 1994). Použitelných molekulárních znaků je celá řada. Znaky pro fylogenetickou analýzu by měly splňovat dvě důležité podmínky. Měly být vzájemně nezávislé (jinak je při analýze nutno brát v úvahu kovarianci, tedy míru závislosti, což komplikuje výpočty) a měly by být vzájemně homologní (Macholán, 2004).

Některé molekulární znaky nevyžadují zjišťování přesné sekvence DNA (či proteinů). Tradiční, dnes již v podstatě historická (Macholán, 2014) je metoda založená na elektroforetické analýze isozymů respektive alozymů. Isozymy jsou funkčně podobné, ale odlišitelné formy enzymů, kódované jedním nebo více geny (v případě, že jsou isozymy produkty různých alel téhož genu, jedná se o alozymy). Jedním z problémů isozymové analýzy je skutečnost, že ne každá změna genu se projeví změnou příslušného proteinu (synonymní mutace způsobené degenerovaným genetickým kódem nejsou pochopitelně elektroforézou proteinů detekovatelné). Navíc lze využít jen geny aktivně exprimující histochemicky detekovatelné proteiny, což je jen malá část genomu zvířete (Park et Moran, 1994). Analýza tzv. mikrosatelitů (SSLP - simple sequence length polymorphism) využívá skutečnost, že se v genomech vyskytují lokusy s mnohonásobnými repeticemi jednoduchého

vzoru. Počet repetic vzoru je variabilní, po amplifikaci pomocí PCR se jeho délka určuje elektroforeticky. V počtu repetic je obvykle velký polymorfismus i v rámci druhu či populace, analýza mikrosatelitů je proto mimořádně výhodná pro vnitropopulační a vnitrodruhové studie (Flegr, 2009).

Pro fylogenetickou analýzu lze využít i štěpení DNA restrikčními endonukleázami (enzymy štěpící v místě specifické krátké sekvence). Restrikční data lze využít buď v podobě kompletní restrikční mapy znázorňující místa štěpení, nebo častěji jen v podobě informací o přítomnosti či nepřítomnosti fragmentů určité délky (RFLP - restriction fragment length polymorphism). Ani RFLP se dnes už pro fylogenetickou analýzu prakticky nepoužívá (Macholán, 2014). Zajímavá je metoda SSCP (single-strand conformational polymorphism). Při té se zkoumá elektroforetická pohyblivost jednořetězcových úseků DNA, vzniklých denaturací původně dvouřetězcové DNA. Jednořetězcová molekula zaujímá, díky vzájemné komplementaritě svých vlastních úseků, trojrozměrnou strukturu, která významně ovlivňuje její pohyblivost. Detekovatelnými změnami mobility se projeví až 90% jednonukleotidových substitucí (Flegr, 2009). Metoda RAPD (randomly amplified polymorphic DNA) využívá amplifikace DNA pomocí krátkých (okolo deseti nukleotidů) primerů, čímž se dosáhne amplifikace mnoha úseků z různých částí genomu. Amplifikované úseky pak vytvoří při elektroforéze komplexní soubor proužků, který muže být druhově specifický. Výhodou je, že nemusíme mít žádnou předběžnou znalost o daném genomu (Macholán, 2014).

Podobnost sekvencí DNA v rámci celého genomu lze srovnávat i hybridizací DNA v roztoku (směs DNA dvou blízce příbuzných druhů renaturuje téměř stejně rychle a dokonale jako v případě DNA z jednoho druhu). Využívat lze i imunologické metody založené na zkřížené reaktivitě protilátek mezi dvěma druhy (Flegr, 2009).

Z hlediska dalšího zpracování jsou ale nejvhodnějším zdrojem molekulárních znaků výsledky sekvenování DNA (Flegr, 2009). Umožňují totiž odhalit nejvíce genetických znaků, kterými se studované druhy liší a umožnují i identifikovat znaky, které jsou pravděpodobně selekčně neutrální (substituce v pseudogenech, intronech a synonymní mutace v oblastech kódujících proteiny) (Flegr, 2009). S rozvojem a zlevnováním metod sekvenování jsou DNA

sekvence stále častějším zdrojem dat pro fylogenetickou analýzu (Flegr, 2009; Macholán, 2014).

Obecně lze molekulární data využitelná ve fylogenetické analýze rozdělit na dvě kategorie. Jednou kategorií jsou data ve formě vzdáleností (distancí, podobností) mezi jednotlivými taxony. Některé výše zmíněné metody, jako hybridizace DNA nebo metody imunologické poskytují data pouze jako vzdálenosti mezi dvojicemi taxonů. Druhou kategorií jsou jednotlivé kvalitativní znaky, jako například restrikční fragmenty nebo nukleotidové sekvence. Důležité je, že kvalitativní znaky lze snadno převést na vzdálenosti mezi taxony, ale opačně to možné není (Avise, 2004).

3.5.3 Metody rekonstrukce fylogeneze

Rekonstrukce fylogeneze je kvalifikovaný odhad skutečné evoluční historie na základě nekompletní informace v dostupných datech, která mohou být i navzájem konfliktní. Výsledkem analýzy je proto hypotéza, kterou je možné porovnávat s jinými hypotézami. Fylogenetické metody lze rozdělit podle dvou kritérií – podle metody konstrukce stromů a podle typu použitých dat (tj. distancí mezi taxony nebo znaků). Podle metody konstrukce stromů lze rozlišit dva principiálně odlišné přístupy, tj. metody algoritmické a metody založené na kritériu optimality (Macholán, 2014).

Algoritmické metody postupují definovanou sekvencí kroků (výpočetním algoritmem) ke konstrukci jediného výsledného stromu. Protože vedou přímočaře ke konečnému řešení, bývají velmi rychlé. Metody založené na kritériu optimality vygenerují všechny přípustné stromy a následně, podle předem zvoleného kritéria optimality, zvolí strom nejlépe vyhovující (nebo více stromů stejně dobře vyhovujících). Protože však musí vygenerovat a vyhodnotit obrovskou množinu stromů, jsou velmi výpočetně náročné (Flegr, 2009). Počet označených binárních stromů s kořenem pro n druhů lze vypočítat podle vzorce:

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

To pro 10 druhů znamená 34459425 stromů a pro 20 druhů již 8200794532637891559375 různých stromů, pro 50 druhů je stromů $2{,}75292 \times 10^{76}$ (Felsenstein, 2004). To vyžaduje značné množství počítačového času. Nevýhodou algoritmických metod je však fakt, že výsledky se mohou lišit podle pořadí v jakém data vstupují do analýzy (Flegr, 2009). Metody založené kritériu optimality jsou matematicky přesnější (Swofford et Sullivan, 2009). Jejich výpočetní náročnost se v praxi snižuje tím, že se nevyhodnocují úplně všechny stromy, ale pomocí různých heuristických metod se některé stromy předem vylučují z analýzy. To však může vést k tomu, že nejlépe vyhovující strom není vůbec nalezen (Flegr, 2009).

Příkladem algoritmických metod založených na distančních datech je metoda UPGMA (nevážená párová metoda aritmetických průměrů) nebo spojování sousedů (neighbor-joining). Mezi metody založené na kritériu optimality a distančních datech patří například metoda minimální evoluce. Příkladem metod založených na kritériu optimality a znakových datech jsou metoda maximální úspornosti (parsimonie), metoda maximální věrohodnosti (maximum likelihood) a Bayesovská analýza (Macholán, 2014).

3.5.3.1 Metoda maximální úspornosti

Metoda maximální úspornosti (parsimonie, maximum parsimony), patřící mezi nejoblíbenější metody konstrukce fylogenetických stromů, je založena na principu preferování jednodušších hypotéz nad složitějšími. To je filosofický princip stojící u základů metodologie vědy (Occamova břitva). Jednoduchostí se v tomto případě myslí co nejmenší počet evolučních kroků. Principem metody maximální úspornosti je vybrat fylogenetický strom s minimální celkovou délkou (to znamená minimálním počtem evolučních událostí) (Macholán, 2014; Felsenstein, 2004).

Využíváme-li metodu maximální úspornosti pro konstrukci fylogenetického stromu na základě sekvencí DNA, nezajímají nás všechny části sekvence. Pozice, které obsahují u všech taxonů stejný nukleotid (tzv. invariabilní místa) jsou z analýzy vyloučena. Ale ani místo, které nese odlišný nukleotid u jediné sekvence, není informativní (výlučně odvozený, autapomorfní znak). Proto je celkový počet potřebných znaků vysoký. Problémem však může být i velký počet homoplázií mezi znaky (tedy situací, kdy se identický znak vyskytuje u různých

organismů, aniž by byl zděděn od společného předka (Flegr, 2009)). V takovém případě nemusí být výsledek spolehlivý ani při velkém počtu použitých znaků (Macholán, 2014).

Existuje několik různých variant metody maximální úspornosti podle toho s jakými vstupními předpoklady pracují. Pro všechny je společná snaha o minimalizaci evolučních kroků. Fitchova parsimonie neuvažuje žádná omezení typů změn, které v evoluci probíhají. Změny mohou probíhat všemi směry a každý znak se může přímo měnit ve kterýkoliv jiný (tedy například A může být substituován C, G i T). Wagnerova parsimonie uvažuje znaky s definovaným pořadím hodnot, kdy změna musí probíhat vždy postupně přes všechny mezihodnoty, pravděpodobnost změny oběma směry je však stále stejná (Macholán, 2004).

Dollova parsimonie je založena na Dollově zákonu o nevratnosti evolučního vývoje, který v jedné ze svých forem říká, že komplexní znak, který jednou vzniknul, nemůže ve stejné formě vzniknout znovu (z tohoto zákona existuje řada výjimek a má i řadu různých formulací) (Felsenstein, 2004).

Dollova parsimonie předpokládá, že znak existuje ve dvou formách, ancestrální formě (0) a komplexní odvozené formě (1). Forma 0 se může změnit ve formu 1 pouze jednou, ale forma 1 může mnohokrát revertovat zpět ve formu 0. Dollova parsimonie se snaží minimalizovat počet těchto reverzí (Felsenstein, 2004). Dollova parsimonie se používá například jako hrubý model v případě práce s restrikčními daty (ztráta restrikčního místa je pravděpodobnější než jeho nabytí) nebo s introny (stejný intron není téměř nikdy začleněn na stejné místo genomu). Dollova parsimonie ale může vést k velkému nadhodnocení nutných evolučních kroků, pokud skutečně daný komplexní znak vznikne vícekrát. Proto se například používá modifikace s tzv. uvolněným Dollovým kritériem, kdy například dvojnásobný vznik znaku upřednostníme před jedním vznikem a deseti ztrátami (Macholán, 2014). Naopak Camin-Sokalova parsimonie předpokládá možnost změny znaku v jiný, nikoliv však změnu v opačném směru (reverzi). To odpovídá malým delecím DNA. V případě větších, překrývajících se delecí, již metoda vhodná není (Felsenstein, 2014).

Vážená (transverzní) parsimonie přiřazuje větší váhu transverzím než transicím. Transverze jsou totiž vzácnější a tak nehrozí tak vážné nebezpečím, že je v místech

podléhajících rychlé evoluci jedna substituce překryta druhou. Transice mohou být naopak užitečné pro fylogenetickou analýzu mezi blízce příbuznými taxony (Macholán, 2014).

Generalizovaná parsimonie představuje zevšeobecnění různých metod tím, že ke každému typu změny přiřadíme určitou hodnotu nákladů na změnu. Výsledkem je matice nákladů (cost matrix) neboli kroková matice (step matrix). Čím je však metoda obecnější, tím vyšší je výpočetní náročnost (Macholán, 2014).

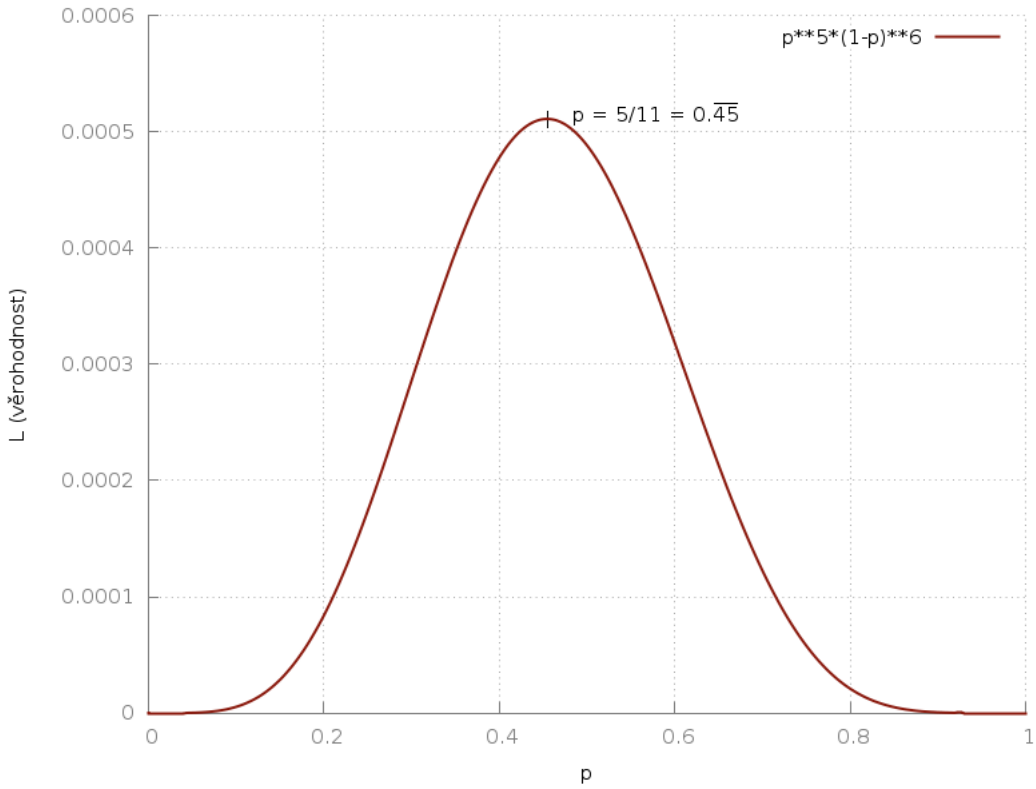
Metoda maximální úspornosti je založená na kritériu optimality a tudíž vyžaduje vyhledání nejlepšího stromu (či stromů) mezi všemi možnými. Prohledat ale skutečně všechny možné stromy je vzhledem k jejich množství velmi výpočetně náročné, zejména v případě dlouhých sekvencí a většího množství taxonů. Proto se v takových případech využívá různých heuristických metod, například postupného přidávání taxonů. Tyto metody výpočet velmi urychlují, ovšem nezaručují nalezení nejlepšího možného stromu (Macholán, 2014).

3.5.3.2 Metoda maximální věrohodnosti

Metoda maximální věrohodnosti (maximum likelihood) je jednou z centrálních metod matematické statistiky a používá se pro řešení řady úloh z různých oblastí. Příkladem může být například hod mincí. Předpokládáme-li, že výsledky jednotlivých hodů mincí jsou navzájem nezávislé, pravděpodobnost panny (P) je vždy p, pravděpodobnost orla (O) je tudíž p-1 a získáme-li jedenácti hody sekvenci PPOOPOPPOOO, lze snadno spočítat pravděpodobnost takového výsledku (Felsenstein, 2004):

$$L = pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p) = p^5(1-p)^6$$

Pro hodnoty pravděpodobnosti od nuly do jedné pak získáme následující křivku hodnot věrohodnosti (vytvořeno v programu gnuplot). Její maximum je při pravděpodobnosti rovné 5/11 (což lze ověřit pomocí derivace). Maximální věrohodnost je tedy při p = 5/11. (Felsenstein, 2004):



Porovnávat různé hypotézy pak lze pak pomocí věrohodnostního poměru (Macholán 2014).

Výpočet věrohodnosti fylogenetického stromu vychází ze seřazených (aligned) DNA sekvencí a z fylogenetického stromu s určitou topologií a délkami větví (jejichž délka vyjadřuje pravděpodobnost změn podél nich, nikoliv čas). Je potřeba evoluční model, který umožnuje vypočítat pravděpodobnosti změn podél větví. Předpokládá se, že evoluce v jednotlivých pozicích sekvence i v jednotlivých liniích v rámci stromu je nezávislá (Macholán 2014; Felsenstein 2004).

Na rozdíl od příkladu s mincemi je ale nalezení nejvěrohodnějšího stromu úkol příliš složitý, než aby mohl být proveden analyticky. Provádí se numericky a vyžaduje velké množství iterací, proto je výpočetně velmi náročný. (Macholán, 2014).

Srovnáme-li metodu maximální věrohodnosti s výpočetně méně náročnou metodou maximální parsimonie, lze konstatovat, že za předpokladu relativně malých evolučních změn

bývají výsledky obou metod podobné. V případě vysoké frekvence substitucí nebo nerovnoměrné evoluce podél jednotlivých větví stromu mohou být výsledky výrazně odlišné. Důležitým rysem metody maximální věrohodnosti ve srovnání například s metodou maximální parsimonie je explicitní specifikace konkrétního evolučního modelu. To má své zastánce i odpůrce. Kritizovanou nevýhodou je vnášení jisté subjektivnosti do analýzy. Vysoká výpočetní náročnost také znamená, že není možné vždy porovnat všechny možné stromy a může tak být nalezeno pouze lokální, nikoliv globální maximum věrohodnosti (Macholán, 2014).

3.5.3.3 Bayesovská analýza

Bayesovská fylogenetická analýza je blízce příbuzná metodě maximální věrohodnosti, ale kombinuje odhad pravděpodobnosti vzniku studovaných dat za předpokladu určité hypotézy s tím, co o datech a priori předpokládáme. Výsledná veličina se nazývá aposteriorní pravděpodobnost a vyjadřuje pravděpodobnost správnosti specifikované hypotézy při existenci získaných dat (Macholán, 2014). Metoda je založena na matematické větě, jejímž autorem je anglický matematik Thomas Bayes:

$$P\left(H\middle|D\right)=\frac{P\left(D\middle|H\right)P\left(H\right)}{P\left(D\right)}$$

P(D|H) je podmíněná pravděpodobnost dat při určité hypotéze (věrohodnost), P(H) je apriorní pravděpodobnost hypotézy a P(D) je suma čitatelů pro všechny možné hypotézy H (Macholán, 2014).

V obecné rovině lze Bayesovu větu demonstrovat na hracích kostkách. Vybíráme hrací kostku ze sady, kde je přesně 20% falešných kostech (na kterých častěji padá šestka), což víme (a priori pravděpodobnost). Hodíme dvě čísla, padne dvojka a šestka. Pravděpodobnost takového výsledku je ale vyšší v případě falešné kostky (18/441) než kostky pravé (1/36). Dosadíme-li tyto hodnoty do Bayesovy rovnice, vyjde nám aposteriorní pravděpodobnost 0,269. Odhad pravděpodobnosti, že je kostka falešná, se tedy na základě výsledků dvou hodů zvýšil (Macholán, 2014).

Důležitým faktorem Bayesovské analýzy jsou proto apriorní pravděpodobnosti, které musí být zvoleny. Narozdíl od příkladu s kostkami je totiž ve fylogenetické analýze zpravidla neznáme. Metody pro volbu apriorních pravděpodobností jsou různé, včetně výběru z arbitrárního rozdělení pro délky větví (např. exponenciální) či uniformního rozdělení (flat priors) (Macholán, 2014). Je nutné si uvědomit, že apriorní pravděpodobnosti je nutné vždy zvolit, i pokud necháme výběr na použitém programu, provedl výběr vlastně autor softwaru (Felsenstein, 2004). Ohledně apriorních pravděpodobností a jejich volby panují jisté kontroverze. Obvykle ale nejsou výsledné aposteriorní pravděpodobnosti na změny apriorních pravděpodobností příliš citlivé a tato citlivost klesá s rostoucím množstvím dat (Macholán, 2014).

3.5.4 Software pro fylogenetickou analýzu

PHYLIP (the PHYLogeny Inference Package) je balík fylogenetických programů, jehož autorem je Joseph Felsenstein. Původní verze v Pascalu je z roku 1980, od roku 1993 je psán v jazyce C (Felsenstein, 2004). Je k dispozici zdarma pod opensource licencí na adrese http://evolution.gs.washington.edu/phylip.html včetně dokumentace. Jsou k dispozici i spustitelné soubory pro Windows, Mac OS X i Linux. Obsahuje řadu dílčích programů, včetně programů pro maximální parsimonii, maximální věrohodnost a další. Programy se ovládají pomocí textového menu a data se zadávají v textovém souboru. Generované stromy jsou v textových souborech ve standardním Newick formátu.

Například program Dnaml z baláku PHYLIP slouží k tvorbě stromů na základě metody maximální věrohodnosti. K jednoduchému otestování funkce jsem použil 5 náhodně vybraných sekvencí fragmentů mitochondriálního genu COI labyrintních a příbuzných ryb stažených ve formátu FASTA z databáze BOLD (BLB008-10 *Channa striata*, ANGEN114-15 *Trichogaster fasciata*, GBGCA9560-15 *Macropodus opercularis*, OFBI073-11 *Trichogaster lalius* a DSANA076-08 *Betta splendens*). Alignment sekvencí byl proveden programem Clustal Omega (webové rozhraní, http://www.ebi.ac.uk/Tools/msa/clustalo/). Výsledek byl ve formátu PHYLIP předán ke zpracování programu Dnaml:

```
Nucleic acid sequence Maximum Likelihood method, version 3.696


Empirical Base Frequencies:

   A       0.25594
   C       0.24140
   G       0.17089
  T(U)     0.33177


Transition/transversion ratio =   2.000000


     +-------Betta sple
  +--1
  |  +-----Macropodus
  |
  |  +----Trich. lal
  2--3
  |  +-----Trich. fas
  |
  +-------Channa str


remember: this is an unrooted tree!

Ln Likelihood = -3605.62215

 Between          And            Length         Approx. Confidence Limits
 -------          ---            ------         ------- ---------- ------

   2       Channa str     0.13433        (  0.09815,    0.17051) **
   2        1             0.04778        (  0.02272,    0.07284) **
   1       Betta sple     0.14117        (  0.10549,    0.17685) **
   1       Macropodus     0.09439        (  0.06424,    0.12454) **
   2        3             0.05420        (  0.02904,    0.07936) **
   3       Trich. lal     0.08420        (  0.05599,    0.11240) **
   3       Trich. fas     0.09691        (  0.06645,    0.12738) **

   *  = significantly positive, P < 0.05
   ** = significantly positive, P < 0.01
```

Přestože se jedná pouze o jednoduchý test programu (jeden úsek genu nepostačuje pro kvalitní analýzu), vytvořený fylogenetický strom odpovídá současným poznatkům o

fylogenezi labyrintek a jejich taxomii – *Trichogaster* je jeden rod, *Betta* a *Macropodus* patří do jedné podčeledi a *Channa* je z úplně jiné čeledi, mimo vlastní labyrintní ryby (Rüber et al., 2006).

PAUP* (http://paup.csit.fsu.edu/) je dalším balíkem programů pro fylogenetickou analýzu. Název původně znamená Phylogenetic Analysis Using Parsimony, ale dnes podporuje i další metody, včetně maximální věrohodnosti (Swofford, n.d.). Narozdíl od balíku PHYLIP ale není zdarma k dispozici. Podobně jako PHYLIP patří k balíkům nejpoužívanějším a nejcitovanějším v odborné literatuře (Felsenstein, 2004).

Pro bayesiánskou fylogenetickou analýzu se používá program MRBAYES. Protože řešení nelze nalézt analyticky, používá MRBAYES pro aproximaci aposteriorních pravděpodobností metodu Monte Carlo Markovových řetezců (MCMC, Markov chain Monte Carlo) a její modifikaci MCMCMC (Metropolis-coupled MCMC). Monte Carlo jsou metody, které využívají náhodného výběru vzorků, přičemž s velkým množstvím náhodných se výsledný fylogenetický strom blíží skutečnému (Macholán, 2014). Program je pod svobodnou licencí a v současnosti dostupný včetně zdrojových kódu na serveru SourceForge (http://mrbayes.sourceforge.net/).

3.5.5 Příklady fylogenetických analýz

V zajímavé studii byl použit okoun říční (*Perca fluviatilis*) pro výzkum tras, kterými sladkovodní ryby rekolonizovali postglaciální Evropu. Okoun je vhodný mimo jiné proto, že snáší vysoký rozsah teplot, což umožňuje časnou rekolonizaci, a také není, narozdíl třeba od lososovitých ryb, šířen člověkem. Využito bylo sekvenování části mitochondriální DNA (D-smyčka) a pro doplnění i metoda RAPD. Analyzováno bylo 55 evropských a jedna sibiřská populace. Identifikováno bylo 35 mitochondriálních haplotypů a z nich byl vytvořen kladogram. Následně byl porovnán s geografickým rozšířením haplotypů. Data podporují hypotézu, že současné západoevropské a severoevropské byly kolonizovány ze třech hlavních refugií v jihovýchodní, severovýchodní a západní Evropě (Nesbø et al., 1999).

Gerlach et al. (2001) využili pět mikrosatelitovývh lokusů pro analýzu populace okounů říčních v Bodamském jezeru. Zjistili, že v jezeře žijí dvě velké, geneticky odlišné populace,

jedna v západní a druhá ve východní části jezera. Zjistili také, že okouni mají tendenci se sdružovat v hejnech příbuzných jedinců (Gerlach et al., 2001).

3.5.6 Využití masivně paralelního sekvenování pro fylogenetické analýzy

Biogeografie tradičně zkoumala biodiverzitu na základě geografického rozšíření druhů. Molekulární metody přinesly nové možnosti, například porovnávat navzájem biologické a geografické kladogramy a objevovat tak jejich souvislosti, geografické procesy určující rozšíření druhů (viz např. Nesbø et al. (1999) a výzkum postglaciální rekolonizace evropských sladkých vod). Analýzy založené na malém počtu lokusů ale neposkytují všechny potřebné informace. Analýza většího množství nezávislých lokusů umožňuje získat o historii druhů více informací a s větší jistotou. Masivně paralelní sekvenační metody nové generace umožňují získat dostatek sekvenčních dat, důležitou limitací je však stále zpracování tak obrovského množství dat (Rocha et al., 2013). Aby bylo masivně paralelní sekvenování ekonomické, je ale nutné sekvenovat mnoho vzorků najednou (různé vzorky jsou označeny speciálními sekvencemi, tagy, přidanými pomocí PCR nebo ligace, takže je lze při zpracování výsledků odlišit) (McCormack et al., 2013).

V molekulární fylogenetice bylo tradičně využíváno malé množství genů reprezentujících jen malý zlomek celého genomu. Důvod pro použití většího množství nezávislých genů (lokusů) je ten, že stromy vytvořené z jednotlivých genů jsou navzájem často diskordantní (neshodné) (Rocha et al., 2013). Zahrnutím informací o více lokusech do fylogenetocké analýzy lze kompenzovat náhodné odchylky jednotlivých genů (McCormack et al., 2013).

Před vlastním masivně paralelním sekvenováním je třeba DNA připravit tak, aby obsahovala vhodné množství (dostatečně velké, ale ne příliš, abychom dokázali výsledky zpracovat) ortologních (navzájem si odpovídajících) lokusů (McCormack et al., 2013). Existuje několik způsobů jak toho dosáhnout. Volba metody závisí na množství požadovaných lokusů, vypočetní kapacitě atd. (Rocha et al., 2013).

Sekvenování PCR amplikonů (amplicon sequencing) je metoda vhodná pro malé množství lokusů, protože pro každý z lokusů je potřeba samostatná PCR reakce. Výhodou je dobrá reprodukovatelnost díky použití specifických primerů (Rocha et al., 2013).

Metoda zvaná targeted enrichment umožňuje sekvenovat tisíce lokusů bez nutnosti PCR reakcí specifických pro jednotlivé lokusy. Využívá vychytávání specifických DNA sekvencí jejich hybridizací s DNA sondami, které jsou ukotveny například na mikrokuličkách (Rocha et al., 2013).

Některé metody jsou založené na štěpení genomu pomocí restrikčních endonukleáz, například RAD (Restriction-site Associated DNA). K místům rozštěpeným restriktázou je přiligován speciální adaptér, DNA je pak fragmentována na ještě menší části a sekvenovány jsou jen části obsahující adaptér. Tak je sekvenována DNA v blízkosti restrikčních míst (Davey et Blaxter, 2010). Velká variabilita restrikčních míst předurčuje tuto metodu pro analýzu populací či blízce příbuzných druhů (Rocha et al., 2013).

Další možnou metodou je využití transkriptomů (tedy RNA molekul vniklých transkripcí z DNA). Transkriptom je vlastně jakási přirozeně redukovaná forma genomu. Nejprve je extrahována RNA, následně je provedena reverzní transkripce do cDNA a následně DNA sekvenace. Výhodou transkripomu je fakt, jsou analyzovány funkční geny, které mohou být ovlivněné selekcí - to může být využito k odhalení funkce genů a jejich vztahu k prostředí. Nevýhodou transkriptomů je skutečnost, že RNA degraduje mnohem rychleji než DNA (je méně stabilní) a jsou proto vyžadovány tkáňové vzorky vysoké kvality (Rocha et al., 2013).

Poslední možností, která je vzhledem k rozvoji metod sekvenování stále dostupnější, je sekvenování celých genomů. Takové studie jsou však stále omezeny nedostatečným rozvojem bioinformatiky, na hardwarové i softwarové úrovni (Rocha et al., 2013). Nicméně počet plně osekvenovaných rybích genomů neustále roste. Prvním kompletně osekvenovaným genomem obratlovce (s výjimkou člověka) byly genom rybí (čtverzubec fugu *Takifugu rubripes* v roce 2002). V roce 2004 byl osekvenován genom dalšího čtverzubce - čtverzubce zeleného (*Tetraodon nigroviridis*) (Koepfli et al., 2015). Čtverzubci mají totiž velmi malý, kompaktní genom, genom čtverzubce zeleného je nejmenší známý ze všech obratlovců (Jaillon et al., 2004). Brzy byly plně osekvenovány i další ryby, zejména druhy, které slouží jako modelové organismy v genetickém výzkumu - medaka japonská (*Oryzias latipes*), koljuška tříostná (*Gasterosteus aculeatus*), dánio pruhované (*Danio rerio*, známá zebrafish), a plata skvrnitá

(*Xiphophorus maculatus*). K prosinci 2014 byly plně osekvenovány a publikovány 2 druhy kruhoústých (Cyclostomata), 1 druh paryb (Chondrichthyes), 19 druhů paprskoploutvých a 2 druhy nozdratých (Sarcopterygii), řada dalších druhů byla před dokončením či publikací (Koepfli et al., 2015). Projekt Genome 10K si klade za cíl osekvenovat 10000 obratlovčích genomů, z toho má být 4000 genomů rybích (Bernardi et al., 2012).

Využití masivně paralelního sekvenování a velkého množství lokusů může přinést zcela nové poznatky. Například hybridizace mezi mořskými rybami se ukazuje jako mnohem častější než se dříve soudilo a studium umožňuje studovat genový tok během speciace a retikulátní evoluce (při které se linie spojují) (Rocha et al., 2013).

4 Závěr

Molekulární metody a zejména sekvenování DNA zažívají v posledních letech obrovský rozmach. Technologický vývoj v oblasti sekvenovacích metod během posledních deseti let (masivně paralelní metody) sekvenování nesmírně zrychlil, zlevnil a zpřístupnil pro řešení problémů, kde to dříve bylo technicky či ekonomicky nemyslitelné. Rychlost sekvenování se zvyšuje takovou rychlostí, že překonává i rychlost rozvoje výpočetní techniky a zpracování a analýza dat se stává stále významnějším problémem vyžadujícím pokročilé statistické, matematické a algoritmické metody. Rozvoj sekvenování umožnuje využívat sekvenční data ve stále více oblastech, od fylogenetiky, kde probíhá posun od využití jednotlivých genů k využití větších částí genomu, což například umožňuje Identifikace druhů pomocí sekvence DNA již umožnuje pracovat nikoliv se vzorky tkáně, ale i s environmentální DNA, získanou ze vzorků prostředí. A metody sekvenování stále prochází bouřlivým rozvojem.

Domnívám se, že sekvenování DNA bude stále významnějším nástrojem ichtyologie i celé biologie. S tím bude nabývat i význam souvisejích analytických metod a bioinformatiky obecně. Zajímavé je i to, že se zlevňováním a zpřístupňováním molekulárních metod, se jejich základy stávájí dostupné nejen odborné ale i amatérské veřejnosti, především v rámci tzv. hnutí DIYbio. Některé DIYbio skupiny například pořádaly pro veřejnost workshopy DNA barcodingu pro ověření pravosti rybího masa.

5 Seznam literatury

Ahmadian, A., Ehn, M., Hober, S. 2006. Pyrosequencing: history, biochemistry and future. Clinica Chimica Acta. 363 (1-2). 83-94.

Ansorge, W. J. 2009. Next-generation DNA sequencing techniques. New Biotechnology. 25 (4). 195-203.

Avery, O. T., MacLeod, M. C., McCarty, M. 1944. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. Journal of Experimental Medicine. 79 (2). 137–158.

Avise, J. C. 2004. Molecular Markers, Natural History, and Evolution. Sinauer Associates. Sunderland, Massachusetts. p. 684. ISBN: 0-87893-041-8.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W. 2013. GenBank. Nucleic Acids Research. 41. D36-D42.

Bernardi, G., Wiley, E. O., Mansour, H., Miller, M. R., Orti, G., Haussler, D., O'Brien, S. J., Ryder, O. A., Venkatesh, B. 2012. The fishes of Genome 10K. Marine genomics. 7. 3-6.

Bayley, H. 2015. Nanopore sequencing: from imagination to reality. Clinical Chemistry. 61 (1). 25-31.

Breu, H. 2010. A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction. Applied Biosystems Whitepaper.

Brinkman, F. S. L., Leipe, D. D. 2001. Phylogenetic analysis. In: Baxevanis, A. D., Ouellette, B. F. F. (eds.). Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. John Wiley & Sons. New York. ISBN: 0-471-38390-2.

Broughton, R. E., Milam, J. E., Roe, B. A. 2001. The complete sequence of the zebrafish (Danio rerio) mitochondrial genome and evolutionary patterns in vertebrate mitochondrial DNA. Genome Research. 11 (11). 1958–1967.

Carvalho, D. C., Neto, D. A. P., Brasil, B. S. A. F., Oliveira, D. A. A. 2011. DNA barcoding unveils a high rate of mislabeling in a commercialfreshwater catfish from Brazil. Mitochondrial DNA. 22 (S1). 97–105.

Carvalho, G. R., Pitcher, T. J. 1994. Editorial. Reviews in Fish Biology and Fisheries. 4. 269-271.

Collins, R. A., Armstrong, K. F., Holyoake, A. J., Keeling, S. 2013. Something in the water: biosecurity monitoring of ornamental fish imports using environmental DNA. Biological Invasions. 15 (6). 1209-1215.

Collins, R. A., Armstrong, K. F., Meier, R., Yi, Y., Brown, S. D., Cruickshank, R. H., Keeling, S., Johnston, C. 2012. Barcoding and border biosecurity: identifying cyprinid fishes in the aquarium trade. Plos ONE. 7 (1). e28381.

Dahm, R. 2005. Friedrich Miescher and the discovery of DNA. Developmental Biology. 278 (2). 274-288.

Davey, J. W., Blaxter, M. L. 2010. RADSeq: next-generation population genetics. Briefings in functional genomics. 9 (5-6). 416-423.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., Thermes, C. 2014. Ten years of next-generation sequencing technology. Trends in genetics. 30 (9). 418-426.

Dunn, M. R., Szabo, A., McVeagh, M. S., Smith, P. J. 2010. The diet of deepwater sharks and the benefits of using DNA identification of prey. Deep-Sea Research I: Oceanographic Research Papers. 57 (7). 923–930.

Ewing, B., Hillier, L., Wendl, M. C., Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research. 8 (3). 175-185.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates. Sunderland, Massachusetts. p. 664. ISBN: 0-87893-177-5.

Filonzi, L., Chiesa, S., Vaghi, M., Marzano, F. N. 2010. Molecular barcoding reveals mislabelling of commercial fish products in Italy. Food Research International. 43 (5). 1383–1388.

Flegr, J. 2009. Evoluční biologie. Academia. Praha. 569 s. ISBN: 978-80-200-1767-3.

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J., Turner, S. W. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature Methods. 7 (6). 461-465.

Galal-Khallaf, A., Ardura, A., Mohammed-Geba, K., Borrell, Y. J., Garcia-Vazquez, E. 2014. DNA barcoding reveals a high level of mislabeling in Egyptian fish fillets. Food Control. 46. 441–445.

Galimberti, A., De Mattia, F., Losa, A., Bruni, I., Federici, S., Casiraghi, M., Mertellos, S., Labra, M. 2013. Food Research International. 50 (1). 55-63.

Gerlach, G., Schardt, U., Eckmann, R., Meyer, A. 2001. Kin-structured subpopulations in Eurasian perch (Perca fluviatilis L.). Heredity. 86 (2). 213-221.

Hebert, P. D., Cywinska, A., Ball, S. L., deWaard, J. R. 2003. Biological identifications through DNA barcodes. Proceedings. Biological sciences / The Royal Society. 270 (1512). 313-321.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., Zamir, A. 1965. Structure of a Ribonucleic Acid. Science. 147 (3664): 1462-1465.

Hutchison III, C. A. 2007. DNA sequencing: bench to bedside and beyond. Nucleic Acids Research. 35 (18). 6227-6237.

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z., Cattolico, L., Poulain, J., De Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J. P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K. J., McEwan, P., Bosak, S., Kellis, M., Volff, J. N., Guigó, R., Zody, M. C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E. S., Weissenbach, J., Roest Crollius, H. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 431 (7011). 946-957.

Jo, H., Gim, J. A., Jeong, K. S., Kim, H. S., Joo, G. J. 2014. Application of DNA barcoding for identification of freshwater carnivorous fish diets: Is number of prey items dependent on size class for Micropterus salmoides? Ecology and Evolution. 4 (2). 219-229.

Kircher, M., Kelso, J. 2010. High-throughput DNA sequencing-concepts and limitations. Bioessays. 32 (6). 524-536.

Koepfli, K. P., Paten, B., Genome 10K Community of Scientists, O'Brien, S. J. 2015. The Genome 10K Project: a way forward. Annual review of animal biosciences. 3. 57-111.

Ledergerber, C., Dessimoz, C. 2011. Base-calling for next-generation sequencing platforms. Briefings in bioinformatics. 12 (5). 489-497.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M. 2012. Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology. 2012. 1-11.

Macholán, M. 2004. Metody analýzy II: Rekonstrukce fylogeneze. In: Zima, J., Macholán, M., Munclinger, P., Piálek, J. Genetické metody v zoologii. Karolinum. Praha. ISBN: 80-246-0795-6.

Macholán, M. 2014. Základy fylogenetické analýzy. Masarykova univerzita. Brno. 289 s. ISBN: 987-80-210-6363-1.

Margulies, M., Egholm, M., Altman,W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S.

M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., Rothberg, J. M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 437. 376-380.

Maxam, A. M., Gilbert, W. 1977. A new method for sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America. 74 (2). 560-564.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., Brumfield, R. T. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular phylogenetics and evolution. 66 (2). 526-538.

Merriman, B., Ion Torrent R&D Team, Rothberg, J. M. 2012. Progress in ion torrent semiconductor chip based sequencing. Electrophoresis. 33 (23). 3397-3417.

Metzker, M. L. 2010. Sequencing technologies - the next generation. Nature Reviews Genetics. 11 (1). 31-46.

Miller, D. D., Mariani, S. 2010. Smoke, mirrors, and mislabeled cod: poor transparency in the European seafood industry. Frontiers in Ecology and the Environment. 8 (10). 517-521.

Morgan, J. A. T., Macbeth, M., Broderick, D., Whatmore, P., Street, R., Welch, D. J., Ovenden, J. R. 2013. Hybridisation, paternal leakage and mitochondrial DNA linearization in three anomalous fish (Scombridae). Mitochondrion. 13 (6). 852-861.

Nesbø, C. L., Fossheim, T., Vollestad, L. A., Jakobsen, K. S. 1999. Genetic divergence and phylogeographic relationships among european perch (Perca fluviatilis) populations reflect glacial refugia and postglacial colonization. Molecular ecology. 8 (9). 1387-1404.

Park, L. K., Moran, P. 1994. Developments in molecular genetic techniques in fisheries. Reviews in Fish Biology and Fisheries. 4. 272-299.

Ratnasingham, S., Hebert, P. D. 2007. bold: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular ecology notes. 7 (3). 355-364.

Riemann, L., Alfredsson, H., Hansen, M. M., Als, T. D., Nielsen, T. G., Munk, P., Aarestrup, K., Maes, G. E., Sparholt, H., Petersen, M. I., Bachler, M., Castonguay, M. 2010. Qualitative assessment of the diet of European eel larvae in the Sargasso Sea resolved by DNA barcoding. Biology Letters. 6 (6). 819-822.

Rocha, L. A., Bernal, M. A., Gaither, M. R., Alfaro, M. E. 2013. Massively parallel DNA sequencing: the new frontier in biogeography. Frontiers in biogeography. 5 (1). 67-77.

Roopnarine, P. 2006. Today is too soon. California wild. 59 (1). 8-12.

Rüber, L., Britz, R., Zardoya, R. 2006. Molecular phylogenetics and evolutionary diversification of labyrinth fishes (Perciformes: Anabantoidei). Systematic biology. 55 (3). 374-397.

Rusk, N. 2011. Torrents of sequence. Nature Methods. 8 (1). 44-44.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., Marçais, G., Pop, M., Yorke, J. A. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Research. 22 (3). 557-567.

Sanger, F., Coulson, A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology. 94 (3). 441-446.

Sanger, F., Nicklen, S., Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 74 (12). 5463-5467.

Sanger, F., Tuppy, H. 1951. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochemical Journal. 49. 481-490.

Schadt, E. E., Turner, S., Kasarskis, A. 2010. A window into third-generation sequencing. Human Molecular Genetics. 19 (2). R227-R240.

Shendure, J., Ji, H. 2008. Next-generation DNA sequencing. Nature Biotechnology. 26. 1135-1145.

Shendure, J. A., Porreca, G. J., Church, G. M., Gardner, A. F., Hendrickson, C. L., Kieleczawa, J., Slatko, B. E. 2011. Overview of DNA sequencing strategies. Current Protocols in Molecular Biology. 96. 7.1.1–7.1.23.

Stein, L. D. 2010. The case for cloud computing in genome informatics. Genome Biology. 11 (207). 1-7.

Steinke, D., Zemlak, T. S., Hebert, P. D. N. 2009. Barcoding Nemo: DNA-Based Identifications for the Ornamental Fish Trade. Plos ONE. 4(7). e6300.

Swofford, D. L., Sullivan, J. 2009. Phylogeny inference based on parsimony and other methods using Paup*. In: Lemey, P., Salemi, M., Vandamme, A.-M. (eds.). The Phylogenetic

Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing. Cambridge University Press. New York. p. 267-312. ISBN: 978-0-521-87710-7.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular ecology. 21 (8). 2045-2050.

Taylor, H. R., Harris, W. E. 2012. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Molecular ecology resources. 12 (3). 377-388.

Thompson, J. F., Milos, P. M. 2011. The properties and applications of single-molecule DNA sequencing. Genome Biology. 12 (2). 1-10.

Valdez-Moreno, M., Quintal-Lizama, C., Gómez-Lozano, R., García-Rivas, M. dC. 2012. Monitoring an Alien Invasion: DNA Barcoding and the Identification of Lionfish and Their Prey on Coral Reefs of the Mexican Caribbean. PLoS ONE. 7(6). e36636.

Ward, R. D., Hanner, R., Hebert, P. D. 2009. The campaign to DNA barcode all fishes, FISH-BOL. Journal of fish biology. 74 (2). 329-356.

Watson, J. D., Crick, F. H. C. 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature. 171 (4356). 737-738.

Wong, E. H.-K., Hanner, R. H. 2008. DNA barcoding detects market substitution in North American seafood. Food Research International. 41 (8). 828-837.

Wu, R. 1970. Nucleotide sequence analysis of DNA: I. Partial sequence of the cohesive ends of bacteriophage λ and 186 DNA. Journal of Molecular Biology. 51 (3). 501-521.

Wu, R., Kaiser, A. D. 1968. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. Journal of Molecular Biology. 35 (3). 523-537.

Wu, R., Taylor, E. 1971. Nucleotide sequence analysis of DNA: II. Complete nucleotide sequence of the cohesive ends of bacteriophage λ DNA. Journal of Molecular Biology. 57 (3). 491-511.

Zouros, E. 2000. The exceptional mitochondrial DNA system of the mussel family Mytilidae. Genes & Genetic Systems. 75 (6). 313-318.

5.1 Internetové zdroje

Bold Systems. BOLD Identification Engine [online]. Bold Systems. 2014. [cit. 2015-4-10]. Dostupné z <http://www.barcodinglife.org/index.php/resources/handbook?chapter=2_databases.html&section=id_engine>.

GenomeWeb. Helicos BioSciences Files for Chapter 11 Bankruptcy Protection [online]. Genomeweb LLC. 2012. [cit. 2015-4-4]. Dostupné z <https://www.genomeweb.com/sequencing/helicos-biosciences-files-chapter-11-bankruptcy-protection>.

GenomeWeb. Roche Shutting Down 454 Sequencing Business [online]. Genomeweb LLC. 2013. [cit. 2015-4-4]. Dostupné z <https://www.genomeweb.com/sequencing/roche-shutting-down-454-sequencing-business>.

Illumina. Sequencing by Synthesis (SBS) Technology [online]. Illumina. 2015. [cit. 2015-4-4]. Dostupné z <http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.html>.

Nobelprize.org. The Nobel Prize in Chemistry 1980 [online]. Nobel Media. 2014. [cit. 2014-11-11]. Dostupné z <http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1980/>.

Pacific Biosciences. SMRT cells [online]. Pacific Biosciences. 2014. [cit. 2015-4-4]. Dostupné z <http://www.pacificbiosciences.com/products/consumables/SMRT-cells/>.

Swofford, D. PAUP* [online]. Sinauer Associates. Není datováno. [cit. 2015-04-04]. Dostupné z <http://paup.csit.fsu.edu/about.html>.

University of Michigan. Interpretation of Sequencing Chromatograms [online]. Není datováno. [cit. 2015-04-19]. Dostupné z <http://seqcore.brcf.med.umich.edu/doc/dnaseq/interpret.htm>.