

Pokroky matematiky, fyziky a astronomie

Lucie Kárná

Genetický kód aneb Studovala příroda teorii kódů?

Pokroky matematiky, fyziky a astronomie, Vol. 56 (2011), No. 2, 89--98

Persistent URL: <http://dml.cz/dmlcz/141992>

Terms of use:

© Jednota českých matematiků a fyziků, 2011

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.



This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* <http://project.dml.cz>

Genetický kód *aneb* Studovala příroda teorii kódů?

Lucie Kárná, Praha

Člověk, mezi jehož zájmy patří vedle matematiky i biologie, brzy zjistí, že tyto dva obory nejsou od sebe tak vzdálené, jak by se na první pohled mohlo zdát. Především molekulární biologie se zabývá mnoha jevy a strukturami, které jsou zajímavé i z pohledu algebry nebo jiných matematických disciplín (např. kombinatoriky, konečné geometrie, teorie grup).

Jedním z těchto fenoménů je molekulární genetika. *Pokroky matematiky, fyziky a astronomie* se již několikrát věnovaly problematice genetického kódu z pohledu algebry (např. [3], [4], [5]). Tentokrát se na genetický kód podíváme očima matematické disciplíny, která má k němu nejbliže: očima teorie kódování.

1. Něco málo z teorie kódování

Teorie kódování popisuje situaci, kdy určitou informaci — *zprávu* — zapisujeme pomocí posloupnosti znaků z nějaké *kódové abecedy*. Důvody k tomuto procesu mohou být různé. Jedním z nich je nemožnost přenést nebo zaznamenat původní zprávu zvolenými technickými prostředky — například Morseův telegraf neumožňoval přenos mluvené řeči, na papír nelze přímo zaznamenat tóny¹⁾ a podobně. Jindy je zpráva v originálním tvaru zbytečně dlouhá — plechovou tabulí s nápisem „Zákaz vjezdu všech motorových vozidel s výjimkou motocyklů bez postranního vozíku“ bychom asi u silnice viděli ještě méně rádi než běžnou dopravní značku B03a.

V současné době je hlavním důvodem kódování zpráv (zpravidla pomocí nul a jedniček) použití digitálního záznamu a výpočetní techniky. Teorie kódování nalézá využití v telekomunikacích, při zpracování signálů, v přenosové a zabezpečovací technice.

Předností kódování je možnost zajistit ochranu zprávy, ať už před nepovolanými očima (šifrování) — to umějí zařídit *kryptografické kódy*, nebo před poškozením způsobeným vnějšími vlivy. Kódy, které umějí odhalit chyby v přenesené zprávě, se nazývají *detekční kódy*; pokud dokážou nalezené chyby i opravit, jsou to kódy *samoopravné*. Jednoduchým příkladem detekčního kódu je rodné číslo, které je konstruované tak, aby bylo dělitelné jedenáctí²⁾.

¹⁾ Noty už jsou určitý druh kódu.

²⁾ To platí jen pro desetimístná rodná čísla; osoby narozené do roku 1954 mají rodné číslo devítimístné a kontrolní číslice, která zajišťuje dělitelnost jedenácti, jim chybí.

1.1. Základní definice

Při kódování používáme dvě množiny znaků: množina \mathcal{A} — *zdrojová abeceda* — slouží k zápisu původní (nezakódované) zprávy; její prvky nazýváme *zdrojové znaky*. Množina \mathcal{B} , pomocí které zapisujeme zakódovanou zprávu, se nazývá *kódová abeceda* a její prvky jsou *kódové znaky*.

Jako *slovo* v abecedě \mathcal{A} , resp. \mathcal{B} označíme libovolnou konečnou a neprázdnou posloupnost prvků z \mathcal{A} , resp. \mathcal{B} . *Kód* v abecedě \mathcal{B} je libovolná množina \mathbf{K} slov v abecedě \mathcal{B} . Obecně tedy na vlastnosti kódu neklademe žádné požadavky. Slova z kódu \mathbf{K} se nazývají *kódová slova*, ostatní slova v abecedě \mathcal{B} jsou *nekódová*.

Speciální postavení mezi kódy mají *blokové kódy*, jejichž všechna slova mají stejnou délku. Příkladem je ASCII kód pro kódování znaků v počítači nebo genetický kód z kap. 3. Naproti tomu slova Morseovy abecedy (kódy pro jednotlivá písmena) mají různou délku a jedná se tedy o kód, který není blokový.

U některých blokových kódů je možné kódové slovo rozdělit na dvě části: *informační* část (složená z informačních znaků) obsahuje kódovanou informaci a *kontrolní* část (tvořená kontrolními znaky) je informační částí jednoznačně určená. Nejpráhlednější je *systematické kódování*, při kterém se za slovo, které chceme zakódovat (a které bude tvořit informační část kódového slova), přidá určitý počet *kontrolních znaků*.

Existují i blokové kódy, pro které systematické kódování neexistuje. Například v kódu „dva z pěti“, který je definován jako množina všech pětiznakových slov obsahujících dvě jedničky a tři nuly, není možné rozdělit znaky na informační a kontrolní.

Blokový kód, který má délku slova n a ve kterém je k informačních a $n-k$ kontrolních znaků, označujeme jako (n, k) -kód. Podíl k/n se nazývá *informační poměr* kódu a udává nám, jaký je poměr „skutečně užitečné“ informace v kódované zprávě. Pro detailnější informace o kódech odkazujeme např. na [1].

1.2. Detekce chyb

Při přenosu zakódované informace v prostoru (vysílání a přijímání zprávy) nebo v čase (uložení informace na paměťové médium a její čtení po delší době) může být zpráva různými vnějšími vlivy narušena. To se na úrovni jednotlivých znaků projeví buď tak, že jeden či více znaků zprávy chybí, nebo jsou naopak přidány nové znaky, nebo je počet znaků zachován, ale dojde k záměně některého (některých) z nich za jiný. První typ poruchy, při kterém dochází ke změně počtu znaků (skluz synchronizace), nyní necháme stranou a budeme se zabývat pouze chybami v blokovém kódu, při kterých je počet znaků zachován.

Jako *detekční kód* označujeme takový kód, který objevuje chyby při přenosu. Toho docílíme poměrně jednoduchým způsobem: Je-li vysláno kódové slovo u a přijato nějaké slovo v z množiny \mathcal{B}^n , mohou nastat dvě možnosti: buď je přijaté slovo v nekódové; pak je zřejmé, že toto slovo určitě nebylo vysláno a došlo k chybě, která je tímto objevena.

Druhá možnost je, že přijaté slovo v je kódové. To ovšem může odpovídat dvěma různým scénářům: buďto je to právě to slovo, které bylo vysláno (to znamená, že přenos

proběhl bez chyby), nebo bylo vysláno nějaké jiné kódové slovo, než bylo nakonec přijato ($v \neq u$). Protože přijímač nemá žádnou možnost rozoznat, který z těchto dvou scénářů nastal, je poslední eventualita ($v \neq u$) nepříznivá — došlo k chybě, kterou kód neobjevil. Pravděpodobnost takové *neodhalené chyby* je důležitý parametr zejména u kódů používaných v bezpečnostně kritických aplikacích (například řízení dopravy nebo jaderných elektráren).

Základním parametrem, ovlivňujícím schopnost kódu odhalovat chyby, je jeho *minimální vzdálenost*. Definujeme-li *Hammingovu vzdálenost* dvou slov u a v stejné délky jako počet znaků, ve kterých se slova u a v od sebe liší, pak *minimální vzdálenost* kódu je nejmenší Hammingova vzdálenost všech dvojic různých kódových slov. Kód s minimální vzdáleností d objevuje všechny t -násobné chyby (tj. nejvýše t chybných znaků ve slově) pro všechna $t < d$, ale neobjeví všechny d -násobné chyby. Přitom však takový kód může objevit i některé další chyby — které další chyby objeví, závisí na jeho detailnější struktuře.

2. Ještě méně z biochemie

2.1. Bílkoviny

Bílkoviny (proteiny) jsou látky, které tvoří základ všech živých organismů. U živočichů je to zcela zřejmé, protože zde bílkoviny mají i výraznou stavební funkci — jsou převažující složkou svalů, kůže, vaziva, důležitou roli mají i v kostech a dalších typech tkání. Rostliny svá těla staví především z polysacharidů, takže by se mohlo zdát, že pro ně bílkoviny takový význam nemají. Ve skutečnosti je ale každý organismus, od bakterie přes pivoňku až po velrybu, zcela závislý na činnosti enzymů, receptorů a dalších nenápadných, ale nesmírně důležitých bílkovin. Můžeme s mírnou nadsázkou říci, že téměř veškerou práci v buňce vykonávají bílkoviny.

Bílkoviny jsou svojí strukturou biopolymery, tvořené primárně lineárním (tj. nevětveným) řetězcem základních stavebních prvků — aminokyselin, vzájemně spojených chemickou vazbou. Těchto aminokyselin je celkem 20 druhů a jedna typická molekula bílkoviny se může skládat ze 100 až 100 000 jednotlivých aminokyselin. Protože jednotlivé aminokyseliny mají různé biochemické vlastnosti, je pořadí aminokyselin v bílkovině určující pro její vlastnosti i prostorový tvar.

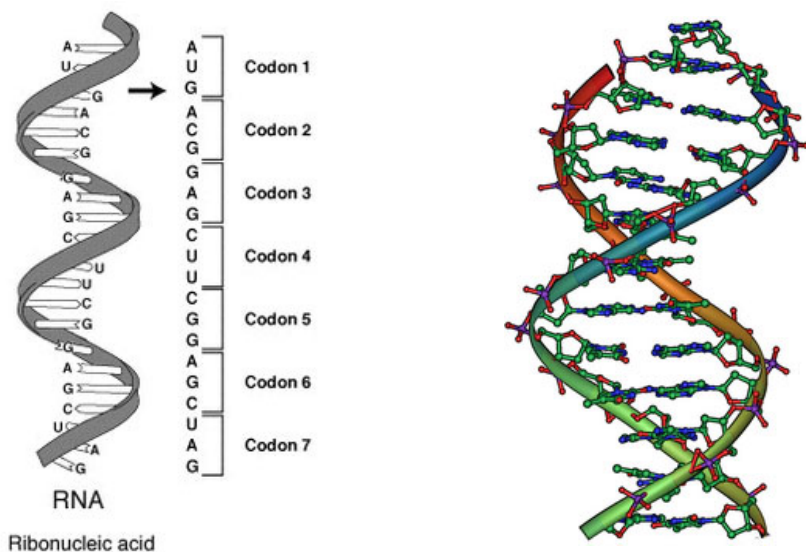
Soubor bílkovin, které má organismus, resp. buňka k dispozici, vymezuje jeho schopnost provádět různé biochemické reakce — zpracovávat chemické látky, získávat z nich energii, vytvářet stavební jednotky svého těla a podobně, a tím do značné míry určuje vlastnosti tohoto organismu.

Proto není nijak překvapivé, že převážná část genetické informace se týká složení bílkovin, a to jejich primární struktury — to znamená pořadí aminokyselin.

2.2. Nukleové kyseliny

Nukleové kyseliny (jmenují se tak proto, že se nacházejí v jádře buňky) jsou informační makromolekuly, jejichž funkcí v organismu je uchování a přenos genetické informace. Jejich dva základní druhy jsou kyselina deoxyribonukleová (DNA), která slouží k dlouhodobému uchování genetické informace, a kyselina ribonukleová (RNA), která hraje důležitou roli především ve zpracování genetické informace a jejím přepisu do struktury bílkovin. Přestože švýcarský lékař Friedrich Miescher objevil nukleové kyseliny již v roce 1871, jejich struktura byla objasněna teprve v roce 1953 pomocí rentgenové krystalografie, viz [6], [7], [8].

Ribonukleová kyselina (RNA) má poněkud jednodušší strukturu než její slavnější sestra. Je to řetězec jednotek, složených z cukru ribózy, nukleové báze a fosfátu, zvaných *nukleotidy*. Nukleotidy jsou lineárně uspořádané a spojené kovalentní chemickou vazbou mezi ribózou a fosfátem; nukleové báze jsou na tomto řetězci přivěšeny stranou. V RNA se vyskytují čtyři druhy bází: adenin (A), guanin (G), cytosin (C) a uracil (U).



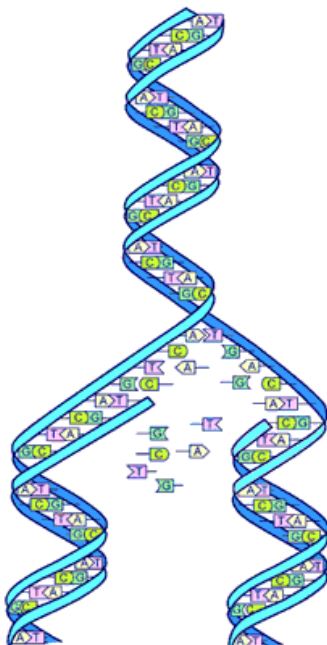
Obr. 1. Nukleové kyseliny: vlevo RNA, vpravo DNA (převzato z [9]).

Podle funkce, kterou v buňce vykonává, se rozlišuje několik typů RNA. V tomto článku se budeme zabývat pouze mRNA (messenger RNA — tedy *poslůček*; někdy se užívá název mediátorová RNA), která slouží jako pracovní kopie genu³⁾. Ostatní typy RNA mají jiné úkoly při syntéze bílkovin — zajišťují transport aminokyselin, tvorbu vazeb mezi nimi, regulaci a podobně.

³⁾ Pod pojmem *gen* se rozumí úsek DNA, který se přepisuje do RNA a následně do konkrétního proteinu.

Deoxyribonukleová kyselina (DNA) má za úkol genetický kód uchovávat dlouhodobě — po celý život organismu a dále ho přenášet do následujících generací. Proto je její struktura stabilnější. Oproti RNA je stavba jejího řetězce mírně pozměněná: jednotlivé nukleotidy jsou chemicky stabilnější (místo ribózy obsahují deoxyribózu) a místo uracilu (U) je zde jiná báze — thymin (T).

Hlavním rozdílem je však její unikátní dvoušroubovicová struktura. Molekulu DNA tvoří dvě paralelní vlákna spojená vodíkovými můstky⁴) mezi nukleovými bázemi do jakéhosi žebříčku, který se přirozeně stáčí do dvoušroubovice. Přitom se spolu vždycky párují jen dvě dvojice nukleotidů: adenin s thyminem (dvěma vodíkovými můstky) a guanin s cytosinem (třemi vodíkovými můstky) — příslušné báze se označují jako *komplementární*. Díky tomu je sekvence nukleotidů na jednom vlákně plně určena sekvencí na druhém vlákně, tvoří vlastně dvojici pozitiv – negativ.



Obr. 2. Replikace DNA (převzato z [9]).

2.3. Tok informace v molekulární genetice

Při dělení buňky je zapotřebí rozdělit i genetickou informaci, a to tak, aby každá ze dvou dceřiných buněk dostala její přesnou kopii. Proces tvorby těchto kopií se nazývá *replikace*. Jeho průběh je zjednodušeně možné popsat tak, že se počínaje určitým pevně daným místem od sebe obě vlákna DNA oddalují a na každém z nich se postupně, nukleotid po nukleotidu, doplňuje nové druhé vlákno. Každá ze dvou nově vzniklých

⁴) Vodíkový můstek je typ slabé vazby; je asi desetkrát slabší než kovalentní chemická vazba.

molekul DNA má tedy jeden řetězec z původní molekuly a druhý nově syntetizovaný; proto se replikace DNA označuje jako *semikonzervativní* proces.

Když buňka potřebuje vyrobit určitou bílkovinu, nejprve vyhledá úsek DNA, který tuto bílkovinu kóduje. Pak se v okolí tohoto genu od sebe oddělí obě vlákna a dvoušroubovice DNA se působením příslušných enzymů částečně „rozmotá“. Tím se stanou nukleotidy na kódujícím vlákně přístupné pro příslušné enzymy, které na ně postupně připojí komplementární nukleotidy RNA a spojí je do řetězce. Nakonec se nově vzniklá mRNA odpojí a obě vlákna DNA se mohou spolu opět spárovat. Tento proces se nazývá *transkripce* neboli přepis (z DNA do RNA). Vzniklá mRNA (*transkript*) slouží jako „pracovní kopie plánu“, u které tolik nevádí, když bude při dalším zpracování poškozena.

Následujícím krokem je vlastní syntéza bílkoviny procesem zvaným *translace* (neboli *překlád*) s použitím mRNA jako plánu. Tento proces přiblížíme v následujících odstavcích.

3. Genetický kód

3.1. Genetický kód v mRNA

Genetický kód je blokový kód, založený na trojicích nukleotidů, přičemž každá trojice (*triplet*, *kodon*) kóduje jednu aminokyselinu. Slovy teorie kódování to znamená, že genetický kód sloužící k syntéze bílkovin podle mRNA (k translaci) je kvaternární blokový kód délky 3 s kódovou abecedou {A,C,G,U}. Jednotlivé triplety jsou jeho kódová slova.

Protože existuje $4^3 = 64$ různých kodonů a jen 20 aminokyselin, zdálo by se, že je zde určitá příležitost k vytvoření detekčního kódu. Ve skutečnosti ale všechny kodony (až na tři) kódují aminokyseliny, takže většina aminokyselin je kódována více než jedním kodonem, tj. genetický kód je *degenerovaný*. Přitom je počet kodonů, které kódují danou aminokyselinu, rozdělen značně nerovnoměrně — zatímco některé aminokyseliny jsou kódovány až šesti různými kodony (arginin, leucin, serin), aminokyseliny metionin a tryptofan jsou kódovány každá jediným kodonem AUG a UGG (viz tabulka na obr. 3)⁵).

Zbylé tři *terminační* kodony nekódují žádnou aminokyselinu, ale definují konec bílkovinného řetězce (ukončují jeho syntézu).

3.2. Detekční vlastnosti genetického kódu

Genetika používá pro chyby v genetickém kódu vlastní terminologii: chyba v genetickém kódu se nazývá *mutace*, jednoduchá chyba (kdy je chybně umístěná jedna báze) je *bodová mutace*.

⁵) Za tabulku genetického kódu získal M. W. Nirenberg v roce 1968 Nobelovu cenu.

	U		C		A		G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	<i>stop</i>	UGA	<i>stop</i>
	UUG	Leu	UCG	Ser	UAG	<i>stop</i>	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Obr. 3. Genetický kód.

Jak je vidět z tabulky genetického kódu na obrázku 3, samotný genetický kód pro translaci (na mRNA) nemá žádnou schopnost detekce chyb. To ani nelze očekávat, protože jeho minimální vzdálenost je pouze 1 (kód obsahuje dvojice kodonů, které se liší jen v jediné bázi a kódují různé aminokyseliny).

Degenerace genetického kódu však přece jen přináší trochu odolnosti vůči chybám (viz [2]). Dva nebo čtyři kodony pro určitou aminokyselinu se typicky liší jen na třetí pozici a záměna v ní tedy sice změní kód, ale nevede ke změně aminokyseliny. Například kodon GGU (kódující aminokyselinu glycin) toleruje jakoukoliv bodovou mutaci na třetí pozici, protože kodony GGC, GGA a GGG také kódují glycin. Takovým mutacím se říká *tiché mutace*.

3.3. Genetický kód v dvouvláknové DNA

Genetický kód v dvouvláknové DNA slouží k dlouhodobému (přes generace) uchování genetické informace. Abychom ho odlišili od genetického kódu v mRNA, který je používán pro translaci, budeme jej nazývat *D-genetický kód*.

Každý nukleotid se páruje se svým protějškem (komplementární bázi) na protějším vlákně. Triplety na dvouvláknové DNA proto můžeme ztotožnit s vektory o 6 nukleotidech.

D-genetický kód na dvouvláknové DNA je tedy kvaternární blokový (6, 3)-kód s kódovou abecedou {A,C,G,T}. Jeho kódové slovo má tři informační znaky na kódujícím vlákně a tři kontrolní znaky na vlákně protějším. Informační poměr D-genetického kódu na DNA je tedy $3/(3+3) = 1/2$.

Protože dvě různá kódová slova D-genetického kódu se liší nejméně dvěma znaky (dvojicí protějších nukleotidů) a existuje dvojice kódových slov, která se liší **právě**

dvěma znaky, je minimální vzdálenost D-genetického kódu rovna 2. Odsud rovnou plyne, že D-genetický kód detekuje všechny jednoduché chyby. Ve skutečnosti je ale schopný odhalit mnohem větší množinu chyb, například všechny dvojnásobné a trojnásobné chyby, které mají všechny chybné znaky na stejném vlákně nebo všechny chyby, kde alespoň jeden pár na protějších pozicích netvoří komplementární báze.

3.4. Samoopravné schopnosti DNA

Transkripce a replikace jsou velmi přesné mechanismy. Při transkripci se objeví přibližně jen jedna chyba za 10^4 bází. Chybovost translace není příliš podstatná, protože jedna vadná molekula bílkoviny obvykle nemůže napáchat velkou škodu. Největší dopad na funkci buňky a jejích potomků může mít chyba při replikaci, proto je její přesnost mnohem vyšší, než je tomu u transkripce.

Již do samotného procesu replikace jsou zabudovány kontrolní a opravné mechanismy. Pokaždé, když je do nově vznikajícího vlákna DNA přidáván nový nukleotid, se kontroluje správnost spárování nukleotidu předcházejícího. Pokud není párování v pořádku, příslušné enzymy chybný nukleotid ihned odstraní a nahradí správným. Díky tomuto kontrolnímu čtení (*proofreading*) je chybovost replikace (pravděpodobnost jednoduché chyby) řádově 10^{-7} . Tak malá chybovost je nesmírně důležitá, vždyť řada zákeřných nemocí (například srpková anémie nebo fenylketonurie) vzniká změnou jen jediného písmene v jediném genu.

Bezprostředně po ukončení replikace části řetězce probíhá další kontrola správného párování bází (*mismatch repair*). Přitom dokáže enzymatický aparát rozeznat původní řetězec DNA od nově syntetizovaného vlákna a chybně spárovaný nukleotid na novém vlákně nahradit správným. V důsledku těchto okamžitých oprav je výsledná chybovost replikace pouhých 10^{-9} .

Mnohem častěji, než k chybě při replikaci, dojde k poškození nukleotidu vnějšími vlivy, jako je například ionizující záření nebo různé silně reaktivní chemické látky (obvykle je nazýváme jedy). Místo jedné ze čtyř základních bází se pak v DNA mohou vyskytnout jejich různé modifikace. Kódová abeceda je tedy vlastně $\{A, C, G, T, Q_1, Q_2, Q_3, \dots\}$, kde Q_i jsou tyto alternativní báze⁶). Protože jsou to evidentně chybné znaky, detekuje genetický kód všechny chyby, kde je alespoň jedno Q_i . Genetický kód na dvouvláknové DNA má navíc i schopnost **opravit** všechny chyby, kde se nevyskytují dvě Q_i proti sobě.

3.5. Mechanismy oprav DNA

Samotná teorie kódování neslibuje žádné samoopravné schopnosti DNA-genetického kódu. Na to, aby byl schopen opravit alespoň jednoduché chyby, by jeho minimální vzdálenost musela být nejméně 3. Příroda si však dokázala poradit jinak. Kromě oprav,

⁶) Například guanin může být změněn na 6-O-metylguanin, který se páruje s thyminem místo s cytosinem.

kteřé probíhají již v průběhu replikace, má buňka k dispozici řadu dalších postupů opravy poškozené DNA. Nejjednodušší možností opravy je přímé nahrazení některých změněných nukleotidů správnými. Opravný mechanismus k tomu nepotřebuje žádný vzor, protože určitá alternativní báze Q_i mohla vzniknout pouze z konkrétní báze správné.

Druhým přístupem je oprava poškozeného úseku dvouvláknové DNA pomocí nepoškozeného druhého vlákna. Enzymatický aparát buňky najde vadný nukleotid (jeden ze znaků Q_i), vystříhne ho i s několika sousedícími nukleotidy, odbourá a chybějící úsek nahradí záplatou syntetizovanou podle nepoškozeného vlákna.

Podobný, ale poněkud složitější je mechanismus *rekombinanatní reparace*, při které proběhne oprava podle sesterské molekuly DNA. (Každá nepohlavní buňka totiž obsahuje dvě kopie každé molekuly DNA — od každého rodiče jednu.)

Bakterie mají jako poslední záchranu ještě jeden mechanismus, takzvanou SOS syntézu. Při ní se silně poškozený úsek DNA (ve kterém mohou být porušena i obě vlákna DNA) přemostí víceméně náhodně zvolenými nukleotidy. Tento proces je pochopitelně značně nepřesný a genetická informace z takto opraveného úseku je pravděpodobně ztracena. Jeho přínos je v tom, že obnoví integritu DNA, která se pak může dále replikovat. S trochou štěstí ztracený úsek nekódoval nic opravdu důležitého, případně jej může bakterie získat od některé sousedky. Tuto schopnost vyměňovat si DNA s jinými jedinci mají pouze bakterie.

3.6. A co člověk?

Velikost lidského genomu (veškeré genetické informace) je přibližně $3,2 \cdot 10^9$ párů nukleotidů. Pokud by se DNA z jediné lidské nepohlavní buňky natáhla a narovнала za sebe, měřila by na délku 2 metry (a na šířku až 2,4 nanometrů).

Četnost chyb v DNA v jedné lidské buňce se odhaduje na 10^3 až 10^6 za den. To se zdá být velmi mnoho, ale je to jen zlomek promile lidského genomu. Mechanismy na opravu DNA přitom pracují velmi usilovně. Pokud jsou u nějakého jedince poškozeny (což se naštěstí stává jen vzácně), strádá některým ze syndromů projevujících se zrychleným stárnutím a úmrtím ve věku odpovídajícím mnohdy věku dětskému.

4. Závěr

Genetické kódy v přírodě vznikaly postupně v průběhu dlouhodobé evoluce trvající již 3,5 miliardy let (tj. cca 10^{17} sekund). V buňce probíhají milióny reakcí každou sekundu, a tak metodou pokusů a omylů dospěla příroda díky Darwinovu přírodnímu výběru až k dnešnímu stavu. Převážná část mechanismů ochrany a oprav genetické informace je založena na biochemických principech, některé lze však — jak jsme viděli — dobře popsat terminologií teorie kódování.

L i t e r a t u r a

- [1] ADÁMEK, J.: *Kódování a teorie informace*. ČVUT Praha, 1991.
- [2] KÁRNÁ, L.: *Genetic code from the Point of View of Code Theory. Proc. Conf. Presentation of Mathematics*, TU Liberec 2006, 267–273.
- [3] KATRNOŠKA, F.: *Genetické algebry*. PMFA 50 (2005), 62–74.
- [4] KATRNOŠKA, F.: *Latinské čtverce a genetický kód*. PMFA 52 (2007), 177–187.
- [5] KATRNOŠKA, F., KRÍŽEK, M.: *Genetický kód a teorie monoidů aneb 50 let od objevu struktury DNA*. PMFA 48 (2003), 207–222.
- [6] WATSON, J. D., CRICK, F. H. C.: *Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid*. Nature 171 (1953), 737–738.
- [7] WATSON, J. D., CRICK, F. H. C.: *Genetical Implications of the Structure of Deoxyribonucleic Acid*. Nature 171 (1953), 964–967.
- [8] WILKINS, M. H. F., STOKES, A. R., WILSON, H. R.: *Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids*. Nature 171 (1953), 738–740.
- [9] Wikimedia Commons: <http://commons.wikimedia.org>.

Proč má DNA tři terminační triplety a jen jeden iniciační?

Michal Křížek, Pavel Křížek, Praha

Stručně z historie kódování DNA

Předchozí znamenitý článek Lucie Kárné [6] nás inspiroval k malému dovětku. Když byla v roce 1953 objevena struktura DNA (viz [11]), začaly prvotní pokusy o určení genetického kódu. Slavného fyzika ruského původu, George Gamowa, uchvátila myšlenka, že k vysvětlení druhové rozmanitosti a fungování genů by mohla být použita kombinatorika a teorie čísel. Jako jeden z prvních si uvědomil, že 20 druhů aminokyselin, z nichž se skládají bílkoviny, nemůže být kódováno dvojicemi nukleotidů, protože existuje jen $16 = 4 \times 4$ různých dvojic nukleotidů ze čtyřprvkové abecedy {A, C, G, T}, kde A je adenin, C cytosin, G guanin a T je thymin. Proto vymýšlel, jak tento nedostatek obejít. Gamow v práci [5] z roku 1954 navrhl tzv. překryvný degenerovaný kód, v němž se sice uvažují jen dvojice nukleotidů, ale příslušná aminokyselina je přiřazena až po přečtení prvního nukleotidu z další dvojice (např. AC GA ...). Je zřejmé, že takový kód (angl. partial overlapping code) by předepisoval velice přísné podmínky na řazení jednotlivých aminokyselin. Později se ukázalo, že tudy cesta nevede.

Prof. RNDr. MICHAL KRÍŽEK, DrSc., Matematický ústav AV ČR, v.v.i., Žitná 25, 115 67 Praha 1, e-mail: krizek@math.cas.cz, Ing. PAVEL KRÍŽEK, Ph.D., 1. lékařská fakulta UK, Albertov 4, 128 00 Praha 2, e-mail: krizekp1@volny.cz