Báze nejsou písmena[1]

Vladimír Matlach — Dan Faltýnek

ABSTRACT:
The Bases Are Not the Letters. In this paper we show some interpretation of the genetic code design. We proceed from the discovery of DNA structure to current stage of the molecular biology. Generally we introduce the basic semiotic assumptions of molecular biology in the description of the structure of DNA, proteins and genetic code. We focus on interpretations of Francis Crick, another molecular biologists, biosemioticians and linguists. For the aims of the paper we describe some fundaments of molecular biology. Core of our text is quantitative analysis (n-gram structure, Zipf's law) of mRNA strings and natural language text. We take into consideration representative quantitative analysis of DNA, RNA and proteins too. Our analysis of mRNA confirms the assumption that the design of the genetic code cannot analogize DNA bases and letters.

KLÍČOVÁ SLOVA / KEY WORDS:
biosémiotika, DNA, genetický kód, jazyková metafora, n-gram, protein, Zipfův zákon
biosemiotics, DNA, genetic code, language metaphor, n-gram, protein, Zipf's law

GENETICKÝ KÓD

Počínaje popsáním struktury DNA mluví biologie, nově vznikající molekulární biologie a obecně věda a společnost o genetickém kódu (viz k tomu Watson — Berry, 2003). Dnes je genetický kód všeobecně intuitivně potvrzovaným vědeckým poznatkem. S předpokladem existence genetického kódu dnes bezprostředně vnímáme živé bytosti a biosféru. Na tento kód se pohlíželo s jistými předpoklady: byly formulovány některé jeho základní vlastnosti a postupem času se pevně ustavilo nahlížení na to, jaký design tento kód má. Výše jsme odkázali k publikaci ozřejmující okolnosti objevu struktury DNA, jejímž autorem je J. Watson. Vlastnosti DNA a genetického kódu ve svých publikacích a přednáškách představoval, vysvětloval a popularizoval Francis Crick. Využijeme jeho textů jako reprezentativních pro představení dnes běžného pohledu na genetický kód. Vlastnosti genetického kódu Crick formuluje následovně.

Genetický kód k sobě vztahuje aminokyseliny (z nichž se skládají proteiny) a báze, které obsahuje DNA (Crick, 1962, s. 11–12; Crick, 1968, s. 367). V procesu výstavby pro-

1 Vladimír Matlach byl podpořen projektem Lingvistická a lexikostatistická analýza ve spolupráci lingvistiky, matematiky, biologie a psychologie, CZ.1.07/2.3.00/20.0161, který je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky. Danu Faltýnkovi bylo zpracování a vydání textu umožněno díky finanční podpoře Filozofické fakulty Univerzity Palackého v rámci Fondu pro podporu vědecké činnosti FF UP.

teinu se k sobě vztahuje 20 aminokyselin a 64 trojkombinací bází, tzv. tripletů (Crick, 1968, s. 368). Tento kód je univerzální a až na výjimky, které ale nevybočují z principu vztahu bází a aminokyselin, je společný všem živým organismům (Crick, 1962, s. 8; Crick, 1968, s. 369).

Báze jsou v genetickém kódu spojeny s aminokyselinami arbitrárním vztahem. Ze všech možných aminokyselin je použito jen konkrétních dvacet, které se vztahují k bázím na základě zprostředkování jistými molekulárními prostředky (tzv. adaptorem — tRNA; Crick, 1967, s. 342–343), přičemž mezi aminokyselinami a bázemi není přímá chemická afinita, mluví se zde o zmrzlé náhodě (Crick, 1968, s. 369). Genetický kód je redundantní, a to v tom smyslu, že více tripletů odpovídá jedné aminokyselině. Některé triplety naopak vztah k aminokyselinám nemají a slouží pouze jako signály vymezující hranici začátku a konce využitelné genetické informace. Relevantní roli v tripletech hrají především první dvě báze. Třetí pozice v tripletu často umožňuje variovat báze, aniž by došlo k záměně aminokyseliny. Umístění bází v tripletu tedy není náhodné. Crick (1968, s. 369) tuto systematičnost detailně popisuje.

Zápis bází je lineární, čte se v jednom směru a bez možnosti přeskakování bází (Crick, 1966, s. 332–333; Crick, 1964, s. 9). Má pevný čtecí rámec, v němž zachovává hranice tripletu. Zároveň se nepřekrývá (z angl. *overlapping*), to znamená, že nenese více informací současně (např. Trifonov ale kód chápe jako překrývající se, důvody popisuje s dalšími spoluautory v práci Popov — Segal — Trifonov, 1996, s. 66; viz k tomu především Trifonov, 1988, s. 508–510). Proces výstavby proteinu, tzv. proteosyntéza, probíhá (informačně a energeticky) směrem od bází k proteinům. Tento princip je nazván centrální dogma.

Uvedený popis genetického kódu byl dále precizován objevy molekulární biologie. Proces proteosyntézy byl popsán s mnoha dalšími proměnami původního řetězce DNA (do hnRNA, mRNA, v souvislosti s interakcemi s snRNA atd.) a procesy směřujícími k finálním produktům proteosyntézy a jejich funkcím. Byly popsány procesy sestřihu, jejich variace mezi organismy a částmi organismů, konformační procesy proteinů, metylace řetězců DNA, chromatinové interakce atd.

Genetický kód je vnímán jako lineární zápis bází vztahující se k tvaru a funkci proteinu. Báze jsou v praxi charakterizovány jako písmena (Crick, 1967, s. 331; Crick, 1962, s. 16). Tato písmena tvoří triplety (trojice bází), které jsou pojímány jako slova, tzv. kodony — kódová slova. Tato slova mají kódem zprostředkovaný vztah ke konkrétní aminokyselině. Soubor tripletů (kódových slov) tvoří gen, jednotku, která má vztah k celému proteinu, jeho tvaru a z něj plynoucí funkci v organismu (Crick, 1962, s. 8; Crick, 1964, s. 2). O bázích jakožto písmenech a tripletech jako slovech se hovoří např. v následujících publikacích (vybíráme reprezentativní doklady pouze pro ilustraci): Stanford (1975, s. 74) ve svých *Základech biofyziky* hovoří o tripletech jakožto o třípísmenných slovech, Weaver (2002, s. 569) ve své *Molekulární biologii* říká, že kodony jsou kódová slova a skládají se ze tří písmen, Twyman (1998, s. 205) ve své syntéze molekulární biologie označuje báze jako písmena, triplety jako slova a geny jako věty a stejně tak se vyjadřují i Hartl a Ruvolová (2013, s. 10).

GENETICKÝ KÓD A PŘIROZENÝ JAZYK

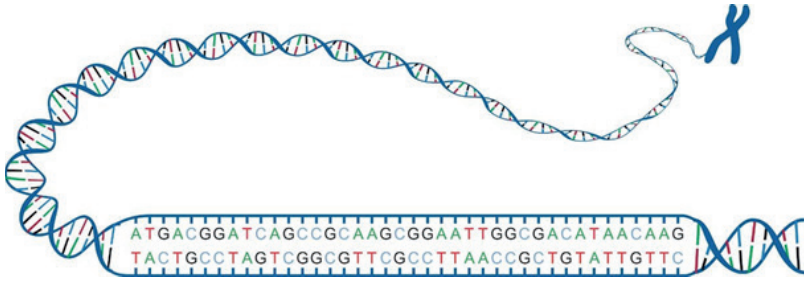
O genetickém kódu se vždy uvažovalo ve vztahu k přirozenému jazyku. Vědní obor molekulární biologie je pevně svázán s pojetím DNA a proteosyntézy jakožto analogie přirozeného jazyka. Analogie molekulárněgenetických procesů s jazykem jsou v molekulární biologii do jisté míry instrumentální, představují tradiční přístup k terminologii. Částečně ale analogie s jazykem molekulární biologii zprostředkovává přístup ke genetickému kódu, jeho struktuře a funkci. Užívání jazykové metafory v molekulární biologii názorně ukázal Raible (2001). Ten provedl korpusové šetření na desítkách tisíc molekulárněgenetických textů, z něhož jasně vyplývá, že termíny analogické k popisu přirozeného jazyka (písmeno, slovo, čtení, zápis, překlad atd.) jsou široce používány v běžné praxi této vědy. Searls (2002) ukázal, že molekulárněgenetický výzkum nesdílí s lingvistikou jen terminologii, ale i výzkumné metody. Jakobson (1971, s. 655–696) dokonce potvrdil molekulární biologii korektnost využívání analogie genetického kódu a jazyka a hovořil přímo o struktuře genetického kódu, který se dle něj skládá z písmen (bází), slov (tripletů) a vět (genů). Dále poukazuje na to, že v genetickém kódu nacházíme vlastnosti, jako je synonymie, suprasegmentální nebo syntaktická delimitace, systém distinktivních rysů či pružná stabilita. Jakobsonův jazykový výklad genetického kódu je pak přejímán dál (např. Katz, 2008). O jazykové metafoře v biologii referují Markoš a Faltýnek (2011).

V tomto textu chceme ukázat, že běžně přijímaný design genetického kódu, jak jsme jej představili výše, je možné zpochybnit. Domníváme se, že pojetí struktury genetického kódu, a to především v jazykové analogii, je od prvopočátku chybné. Chceme jej odmítnout, a to na základě popření analogie bází a písmen. Ze sémiotického hlediska jsme to již udělali (viz Faltýnek, 2012). K tomuto účelu využijeme metodu kvantitativní analýzy textu, kterou představíme níže. Nejdříve ale čtenáře seznámíme se základním instrumentáriem molekulární biologie, které je nutné k vyložení našich závěrů.

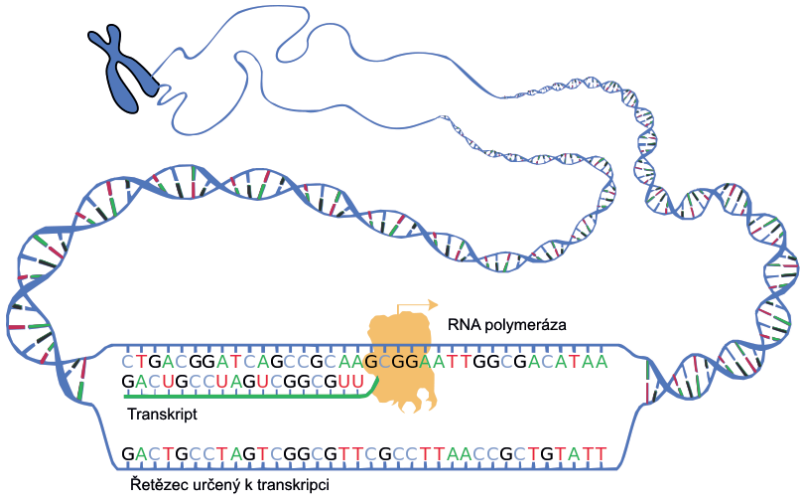
INSTRUMENTÁRIUM MOLEKULÁRNÍ BIOLOGIE

Dědičná informace, obsahující instrukce k výstavbě organismu a řízení jeho biologických pochodů, je fyzicky zapsána v každé buňce ve formě deoxyribonukleové kyseliny mající podobu dvoušroubovice a známé pod zkratkou DNA. Způsob, jakým DNA uchovává informace, je založen na principu variací čtyř specifických makromolekul nukleových kyselin, konkrétně thyminu (T), guaninu (G), adeninu (A) a cytosinu (C). Střídáním těchto tzv. bází dochází k záznamu informace obdobně, jako když Morseova abeceda zaznamenává informace střídáním teček a čárek. Každá z bází má svůj chemicky afinitní (vzájemně vázaný vodíkovými vazbami) protějšek, thymin stojí v DNA vždy proti adeninu a guanin proti cytosinu.

Lineární zápis bází DNA nám dovoluje celou DNA přečíst a přepsat formou textu, tj. zapsat zleva doprava, písmeno za písmenem (viz např. Cvrčková, 2006, s. 17). Takový přepis je kopií řetězce bází zkoumané DNA. Tento postup je v praxi nazýván



OBRÁZEK 1: Dvoušroubovice DNA s jejími dvěma rameny nesoucí jednotlivé báze. A značí adenin, T thymin, G guanin, C cytosin. Dvoušroubovice DNA zaujímá strukturu v tzv. chromozomu (viz: <http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85259>).



OBRÁZEK 2: Transkripce. Přepis sekvence DNA na komplementární protějšky jejích bází RNA polymerázou, thymin je přepisován na uracil, zbylé báze na své komplementární protějšky (viz: <http://commons.wikimedia.org/wiki/File:DNA_transcription.svg>).

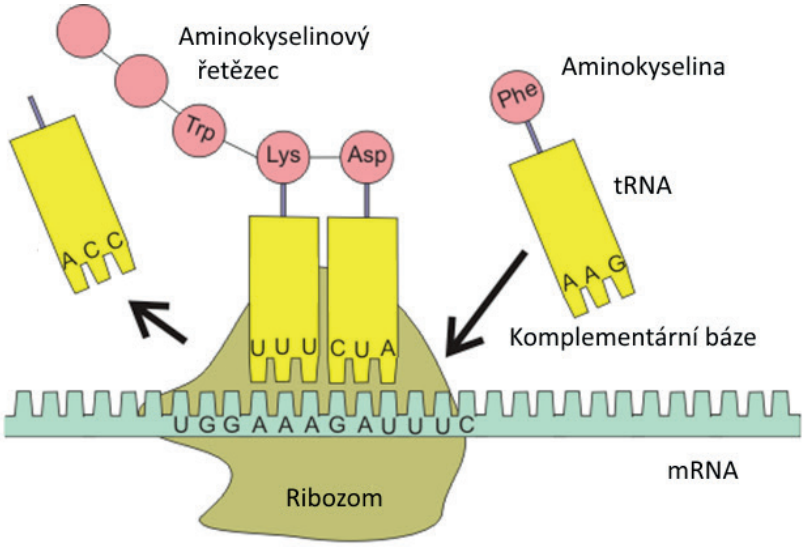
jako sekvenování (detailněji Berg — Tymoczko — Stryer, 2012, s. 140–148). Sekvenováním je řetězec DNA zprostředkován ke zkoumání nejrůznějšími nástroji.

Proces, kterým se z lineární sekvence DNA (tedy z určitého textu) stane protein — reálný fyzický nástroj využitelný v organismu —, označujeme jako proteosyntézu. Celý tento proces můžeme popsat v několika krocích (viz např. Alberts et al., 2008, s. 329; Weaver, 2002, s. 39):

1. Transkripce. Po naplnění specifických podmínek uvnitř buňky se na začátek sekvence DNA (tzv. genu) připne protein RNA-polymeráza, který se po této sekvenci pohybuje v zadaném směru a nukleotid po nukleotidu tuto sekvenci přepisuje. Výsledkem je tzv. transkript — samostatná „pracovní kopie“ DNA ve formě ribonukleové kyseliny RNA, která je určena k zamýšlenému použití v proteosyntéze.

2. Úprava transkriptu. Transkript může být dále upraven (například vystřihnutím částí, které jsou v sekvencích vloženy a slouží jiným účelům). Finální verze transkriptu, tzv. mRNA, je pak přemístěna k ribozomu, kde je konstruován protein.

3. Translace. Ribozom čte mRNA lineárně po trojicích nukleotidů (tripletech). Každému tripletu mRNA je donesena pro něj specifická aminokyselina, která je připojena k předchozí. Takto vytvořený aminokyselinový řetězec se při výstupu z ribozomu začne na základě fyzikálních vlastností jednotlivých konstituentů a fyzikálních vlastností molekul v prostředí (tedy na základě určitého kontextu) formovat, dokud nevytvoří stabilní konformaci (tvar) proteinu. Funkce a vlastnosti proteinu jsou determinovány fyzikálními vlastnostmi jeho makromolekulární konformace (Twyman, 2004, s. 103).



OBRÁZEK 3: Translace. Proces vzniku aminokyselinového řetězce proteinu. Tripletům mRNA je na ribozomu přiřazována tRNA s komplementárním tripletem, která nese aminokyselinu. Takto přiřazené aminokyseliny tvoří řetězec budoucího proteinu (viz: <Boumphreyfr/Wikipedia>).

Tripletům je aminokyselina přiřazena pomocí zprostředkujícího elementu — tzv. adaptorové molekuly tRNA. Adaptor tRNA získává svůj tvar při transkripci z DNA (Weaver, 2002, s. 51). Funkcí této adaptorové molekuly je vázat na jedno ze svých vazebných míst specifický triplet a na své druhé vazebné místo konkrétní aminokyselinu. Vztah aminokyselin a nukleotidů je tak zapsán přímo v DNA. Popsaný vztah byl nazván jako genetický kód (Weaver, 2002, s. 12).

Genetický kód je tvořen variacemi čtyř různých nukleotidů v každé pozici tripletu. Triplet může nabývat 43 (64) možných unikátních kombinací, které tak mohou kódovat 64 různých aminokyselin. Všemi šedesáti čtyřmi realizovanými triplety je kódováno dvacet různých aminokyselin (Alberts et al., 2008, s. 367), mnoho tripletů kóduje stejnou aminokyselinu (kód je tzv. degenerovaný; viz obrázek 4). Některé triplety mají využití jako označení počátku a konce kódující sekvence.

1. pozice	2. pozice U	2. pozice C	2. pozice A	2. pozice G	3. pozice
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	stop	stop	A
	Leu	Ser	stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G
			Aminokyseliny		

OBRÁZEK 4: Genetický kód. Umístění bází v první, druhé a třetí pozici tripletu. Zkratky (Phe, Leu atd.) označují aminokyselinu kódovanou daným tripletem (viz: <www.genome.gov>).

Pojmenování vztahu bází řazených v mRNA a aminokyselin jakožto genetického kódu se vztahuje k jednomu z klasických pojmů lingvistiky. Pojem kód vyjadřuje vztah dvou veličin daný určitým územ (Monod, 1970, s. 159–160). Od pojmenování genetický kód se konzistentně odvíjí i další názvosloví — kromě samotného kódování aminokyseliny tripletem (kodonem) je celý proces tvorby aminokyselinového řetězce podle nukleotidového vzoru při proteosyntéze pojmenován jako translace. Ta se staví do opozice vůči jednoduchému přepisu vzájemně chemicky afinitních molekul při transkripci.

LINGVISTICKÁ PARALELA

V instrumentáriu jsme popsali strukturu DNA a proces proteosyntézy. V souvislosti s tím jsme představili také tradiční molekulárněbiologickou terminologii. Tato terminologie často využívá lingvistických termínů (transkripce, translace, kód, text) nebo termínů založených na jazykové metafoře (zápis, čtení, zpráva, informace (neterminologicky)). Jakobson (1971) potvrzuje, že využití jazykové metafory v molekulární biologii je korektní a že genetický kód má vlastnosti přirozeného jazyka. Analogizuje báze s písmeny (respektive fonémy), triplety se slovy a geny s větami. Nachází

i mnohé další společné vlastnosti genetického kódu a přirozeného jazyka. Ji (1999, s. 412) postupuje dále a analogizuje širokou škálu vlastností genetického kódu a přirozeného jazyka: písmena s nukleotidy a aminokyselinami, slova s geny, řetězce slov se souborem společně exprimovaných genů. Dále k sobě vztahuje gramatiku a fyzikální a chemické zákony, fonetiku a řízení energetického toku, sémantiku a genově řízené procesy v buňce. V případě obou kódů explicitně hovoří o dvojí artikulaci. Ji v nalézání protějšků procesů v buňce a konceptů popisujících přirozený jazyk představuje extrémní případ. Jeho přístup sugeruje, že libovolnému lingvistickému konceptu lze nalézt odpovídající proces či strukturu v buňce.

Trifonov (1988) popisuje soustavy kódů zajišťujících interakci DNA, RNA a proteinů, u jiných autorů můžeme nalézt další obdobné metafory (viz např. Barbieri, 2002; Collado-Vides, 1992; 1993; Markoš, 1997; 2002).

Diskurz těchto znakových popisů procesů v buňce je rozvíjen biosémiotikou, mladou vědní disciplínou. Problém současné biosémiotiky spočívá v tom, že výše zmíněné a mnohé další znakové přístupy k buňce jsou vzájemně nekonzistentní. Každý autor rozvíjí specifický přístup a neexistuje jednotná metoda, která by platnost těchto přístupů ověřovala. Na Jiho příkladu lze vidět, že analogizovat přirozený jazyk a genetický kód lze libovolně, přičemž posouzení korektnosti takových analogií není snadné.

MOTIVACE

Metafory a analogie nám mohou poskytnout nadhled nad určitou problematikou. To je ale v kontextu jazykových metafor a analogií DNA problematické. Nevíme, které z těchto metafor jsou relevantní a užitečné a které nikoliv. Popis procesů v buňce využívá jazykovou metaforu a analogii. Uvažování genetiků, bioinformatiků, makromolekulárních biologů, biochemiků a dalších tak může být ovlivněno zavádějící metaforou. Z epistemologického hlediska by korekce těchto metafor měla pro jejich uživatele velký význam. Cílem tohoto článku je představit experimentální metodu, která by mohla ověření některých ze zmíněných metafor umožnit a zároveň poskytnout vhled do struktury genetického kódu.

METODIKA KVANTITATIVNÍ ANALÝZY DNA

Pro analýzu nukleotidových sekvencí reprezentovaných zápisem v textu jsou užívány různé lingvistické metody a kvantitativní analytické přístupy. Představíme některé z nich a ukážeme, jak se vztahují k naší metodě analýzy struktury genetického kódu. Pokusíme se o využití těchto metod pro podložení nového designu genetického kódu.

Mantegna et al. (1995; viz též Havlin et al., 1995) analyzovali projevy Zipfova zákona na kódující a nekódující DNA. Kódující DNA dle Mantegni et al. Zipfův zákon vykazuje. Nekódující DNA projevy Zipfova zákona vykazuje také, ale pouze do určité míry. Mantegnova analýza byla motivována poznatkem, že pouze malá část (pro

homo sapiens uvažováno 5,33 %; Mantegna et al., 1995, s. 2940) genomu je kódující, a tedy nese informaci k výstavbě proteinu (viz naše instrumentárium výše). Zbylá část genomu nemá takovou jasně zadanou funkci a od šedesátých let se pro ni zažil termín junk DNA. Nekódující DNA měla být v genomu historicky neseným reliktem bez využití v organismu (viz např. Watson — Berry, 2003, s. 253; Palazzo — Gregory, 2014). Mantegna et al. (1995) o junk DNA píše jako o silent DNA.

Mantegna et al. (1995, s. 2949) dále tvrdí, že se nekódující DNA podobá v některých vlastnostech přirozenému jazyku (viz též Niyogi — Berwick, 1995). Mantegna et al. (1995) mluví o tom, že nekódující DNA nese určitý jazyk, z hlediska jeho redundance oproti kódující DNA dokonce bližší přirozeným jazykům (tím Mantegna et al. rozšířili analogie DNA a přirozeného jazyka, o nichž hovoříme níže, a opět uplatnili jazykovou metaforu DNA). Tato zjištění Mantegnu et al. vedou k hypotéze, že nekódující DNA má také funkci, kterou prozatím neregistrujeme a nepopisujeme, a tedy že nekódující DNA je nějakým způsobem použita pro uchování informací „biologických struktur" (Mantegna et al., 1995, s. 2949). Pozdější rozvoj molekulární biologie dal Mantegnově domněnce za pravdu (viz Alberts et al., 2008, s. 31–42; The ENCODE Project Consortium, 2012, s. 57).

Potvrzení výskytu Zipfova zákona u kódující DNA odpovídalo tomu, že kódující řetězec nese informaci k výstavbě funkčního tvaru proteinu, tj. určité struktury s určitou funkcí v organismu. Analogicky k tomu se v textech v přirozeném jazyce projevuje Zipfův zákon z důvodů naplňování určité funkce textu (to se můžeme pokusit vysvětlit např. v souvislosti s informační strukturou textu zajišťující přenos signálu prostředím a výrazovou a obsahovou strukturou a soudržností textu; viz Zipf, 1949, s. 19–47).

K Mantegnově et al. analýze je ale nutné poznamenat následující: Ve své analýze Mantegna et al. využívají dlouhé řetězce nekódující DNA (delší než 50 tisíc bází). Nekódující DNA sice informaci nese, ale nese také množství reliktních řetězců bez využití (nekódují protein ani se nepodílejí na regulaci proteosyntézy). Projekt ENCODE (2012) odhaduje, že až 80 % nekódující DNA má funkční využití. Zbylých 20 % muselo Mantegnovu et al. analýzu ovlivnit, a to proto, že v jeho analyzované nekódující DNA musely být obsaženy složky regulace proteosyntézy a také reliktní DNA (pro niž můžeme stále používat termín junk DNA a která obsahuje mnoho repetitivních a z informačního hlediska redundantních sekvencí). Tato kontaminace by pak posilovala hodnocení nekódující DNA jako podobné přirozenému jazyku z hlediska redundance.

Mantegna et al. byli při zacházení s nekódující DNA postaveni před následující problém: koncept genetického kódu přisuzuje kódující DNA strukturní roli tripletů (viz vztah tripletů a aminokyselin popsaný výše). Nekódující DNA však takové striktní ohraničení a priori přisuzovat nelze, funkční roli zde mohou zastávat sekvence o různé délce. Proto se pro kvantitativní analýzu kódující i nekódující DNA rozhodli využít techniku tzv. n-gramů (tuto techniku využívají např. také Bolshoy et al., 2010, s. 26). Představíme ji na ilustračním příkladu.

Mějme následující řetězec ABCDEFGH. Tento řetězec segmentujeme na 3-gramy, jimiž jsou: ABC, BCD, CDE, DEF, EFG, FGH; 4-gramy mají podobu: ABCD, BCDE, CDEF, DEFG, EFGH. N-gramová analýza tedy segmentuje řetězec tak, že postupuje lineárně jednotku po jednotce a delimituje vždy v řetězci následující n-tici (n-gram). V analýze přirozených textů n-gramová segmentace postupuje bez registrace hranic slov či vět.

Sekvence „Zipfův zákon" je rozdělena na tyto 5-gramy hlásek: zipfů, ipfův, pfůvz, fůvzá, ůvzák atd. Při analýze kódující i nekódující DNA jsou podobně zanedbány jakékoliv dříve stanovené hranice. N-gramový přístup tak dovoluje analyzovat řetězec nezávisle na jeho vnitřní strukturaci tím, že registruje jednotlivé sousedící prvky. Jednotlivé n-gramy představují v analýze analogie slov, která nejsou vydělena mezerou, ale hranicí délky n-gramu (zipfů, ipfův, pfůvz atd. představují slova vstupující do analýzy).

Tento přístup zvolili k oběma typům DNA Mantegna et al. My tímto způsobem postupujeme také, a to z toho důvodu, abychom se vyhnuli apriornímu určení hranic kódovaných složek řetězce (viz dále). Jednotky delimitované n-gramovou technikou budeme stejně jako Mantegna et al. analyzovat z hlediska projevů Zipfova zákona. Výsledky analýzy nás mají vést k potvrzení aktuálního designu genetického kódu, nebo případně k jeho odmítnutí a následné reformulaci.

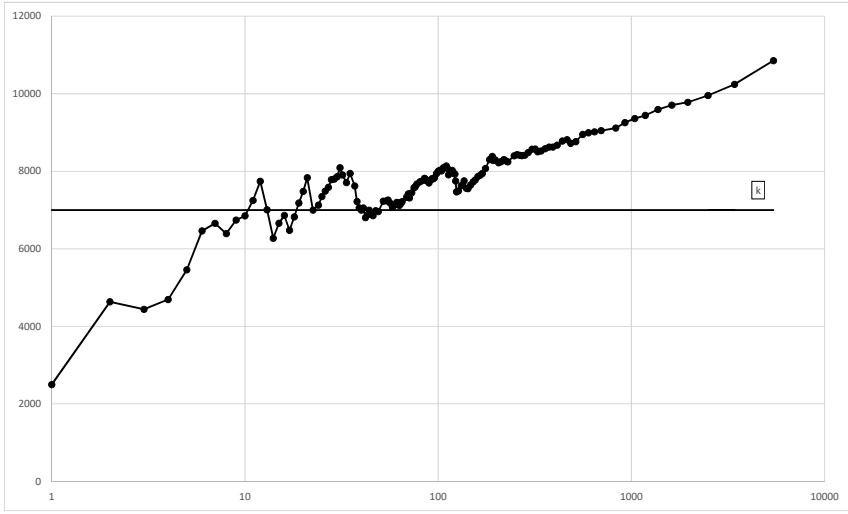
Je však nutné poznamenat, že využití Zipfova zákona v naší analýze může být sporné. Kvantitativní charakteristiky textu, jako jsou projevy Zipfova zákona nebo n-gramové analýzy, mohou být jako důkaz o určité vlastnosti či povaze tohoto textu (je kódující / nese informaci / nese instrukci; znaková funkce v genetickém kódu) chápány jako nepřímé. Např. Konopka (1995) referuje o závěrech Mantegnova týmu a doplňuje poznámky, které mohou být analogicky vztaženy i k našim závěrům. V prvním případě vyvstává problém s tím, že projevy Zipfova zákona jsou identifikovány na mnoha jevech, jako je trh, velikost měst nebo biologických populací atd. Zipfův zákon se pak jeví jako epifenomén jakéhokoliv systémového chování fyzikálních, sociálních, biologických apod. soustav. Zipfův zákon se tedy nemusí vztahovat ke kódujícím funkcím řetězce, ale naopak k jiným jevům jeho konstrukce. V případě řetězce bází by to mohla být jejich kombinatorika daná termostabilitou jednotlivých bází. V případě přirozených textů by se mohlo jednat např. o fonetické důvody kombinovatelnosti hlásek, které představují typ systematizace.

Zipfův zákon byl dokonce napadán v obecném rozměru, a to na základě jeho projevů v náhodných a nenáhodných textech (Li, 1992). Diskuse v této oblasti zahrnuje problém generování náhodného vzorku a to, že mechanismus tvorby náhodného vzorku může produkovat projevy, které se z důvodu ne zcela náhodného generování textu přiblíží Zipfovu zákonu — to by opět potvrzovalo domněnku, že Zipfův zákon je projevem libovolného systémového, auto/regulovaného chování. Stále je ale vnímán jako projev znakovosti, kódovosti či jazykovosti. Na základě něj se hodnotí např. dorozumívání zvířat nebo struktura textu pacientů s postižením způsobujícím jazykový deficit (Ferrer-i-Cancho, 2006; Ferrer-i-Cancho — McCowan, 2009; Ferrer-i-Cancho — Elvevåg, 2010). I přes veškeré zmíněné výhrady použijeme v naší analýze Zipfův zákon, navážeme tak na kvantitativnělingvistický diskurz ověřování kódové povahy DNA, jako je tomu u Mantegni et al. a dalších. K tématu Zipfova zákona a DNA viz Tsonis, Elsner a Panagiotis (1997).

ZIPFŮV ZÁKON NA PŘIROZENÝCH TEXTECH

Zipfův zákon formuluje následující vztah: vezmeme-li určitý text a seřadíme-li počty výskytů (neboli frekvence) jeho entit (např. slov) od nejvyšší po nejnižší, pak po-

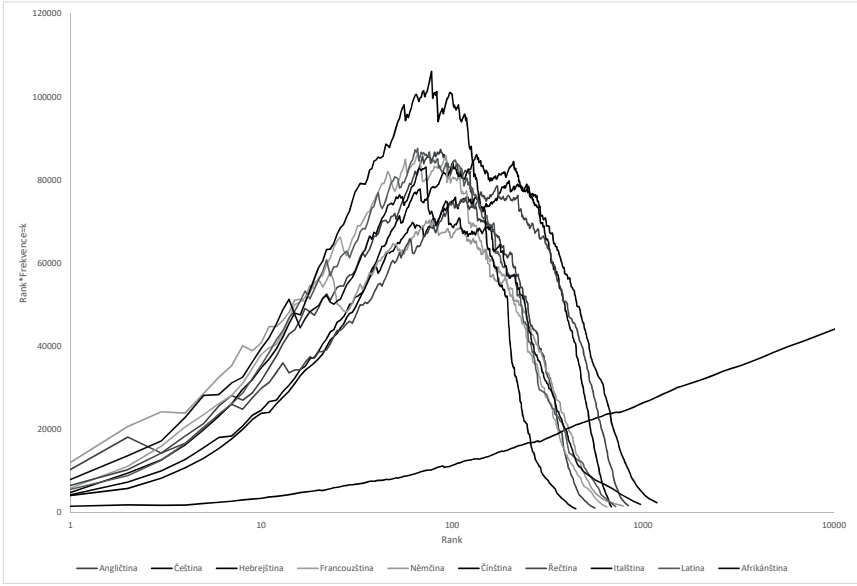
kud vynásobíme frekvenci každé této entity jejím pořadím (tzv. rankem), bude se výsledek *p* vždy blížit určité hodnotě reprezentované tzv. konstantou *k* (Zipf, 1949, s. 22–25). Jak si ale můžeme všimnout na obrázku 5, výsledky násobků ranků a frekvencí jsou u přirozených jazyků značně proměnlivé. Pomocí vodorovné úsečky proto v obrázku ilustrativně zobrazujeme konstantu *k*, jak ji vyjadřuje Zipfův zákon — reálné hodnoty násobků ranků a jejich frekvencí jsou od ní různě vzdáleny.



OBRÁZEK 5: Vztah ranku a násobku ranku a frekvence písmen českého textu a ilustrace konstanty *k* Zipfova zákona. Text: M. Viewegh — Účastníci zájezdu.

Experimentální metoda založená na Zipfově zákonu spočívá ve sledování průběhů grafů hodnot násobků ranků a frekvencí entit textu (tj. spočívá ve vizuální komparaci průběhů grafů a registrování jeho vlastností, např. konkávnost, konvexnost, strmost, charakter maxim, definiční obory, monotónnost atd.; prozatím jsme neaplikovali žádnou formální metodu vyjádření podobnosti grafů, použitá kritéria jsou však pro naše účely dostačující). Tyto entity vybíráme z jedné konkrétní jazykové roviny textu — písmena, slova, věty apod. Klíčovým aspektem této metody a naší experimentálně ověřenou tezí je, že se u různých přirozených jazyků jednotky konkrétních jazykových rovin (písmena, slova, věty apod.) projevují podobně. Máme-li pak text v neznámém jazyce či neznámém zápisu, můžeme díky této metodě identifikovat jazykovou povahu jeho jednotek. Připomínáme jen, že je při této analýze využita n-gramová technika. Příklad uvádíme na obrázku 6. Na něm můžeme sledovat projevy Zipfova zákona u deseti různých jazyků. Studovanými jednotkami jsou zde dvojice písmen textu (2-gramy; nejsou registrovány žádné spřežky, pracuje se s nimi jako s kombinací jednotlivých písmen, např. spřežka *ch*; tečky, mezery, pomlčky atd. nejsou registrovány, registrována jsou pouze písmena). Všechny texty byly před analýzou redukovány na stejný počet znaků (270 000). Čeština, angličtina, němčina, afrikánština, latina, italština, francouzština, řečtina a hebrejština mají obdobný průběh grafu, čínština se od ostatních průběhů výrazně liší. Z tohoto pozorování můžeme

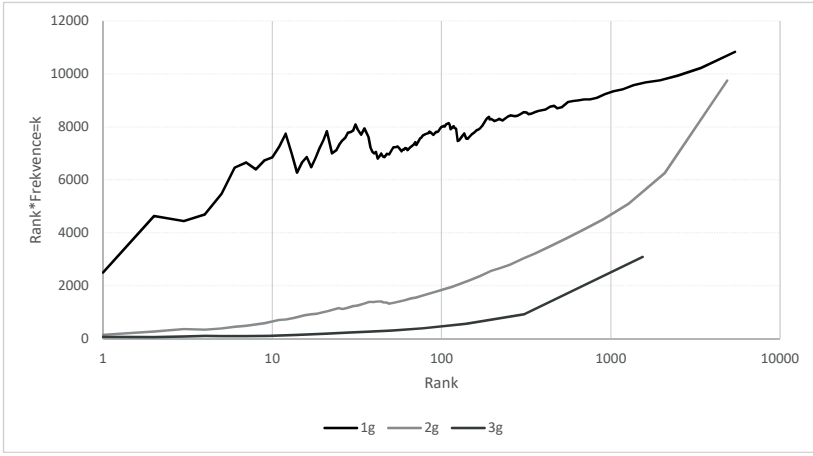
usuzovat, že znaky čínštiny mají oproti ostatním zkoumaným jazykům zcela jinou jazykovou roli.



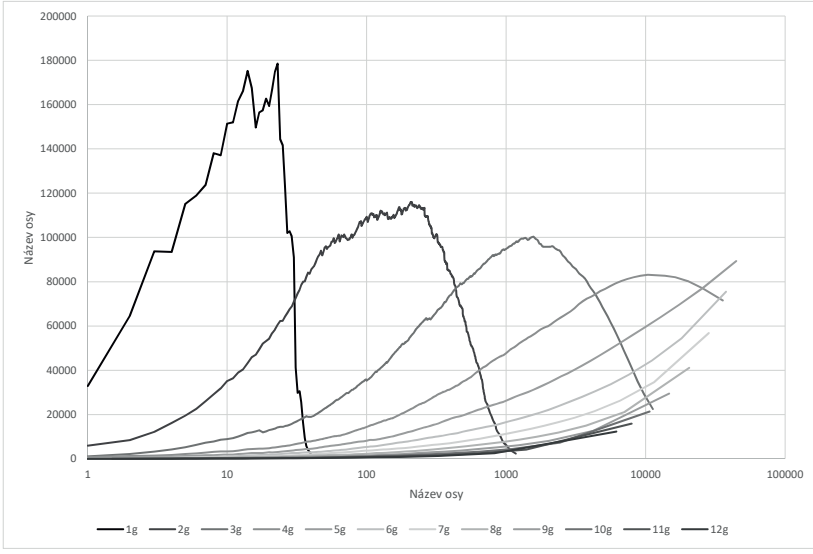
OBRÁZEK 6: Projev Zipfova zákona na 2-gramech písmen textů vybraných jazyků. Texty: afrikánština <18203-8>; řečtina <28658-0>; hebrejština <8cewa10>; latina <27219-0>; němčina <30695-0>; italština <28910-8>; francouzština <15371-8>; čínština <25350-0>; angličtina J. R. R. Tolkien — The Lord of the Rings; čeština M. Viewegh — Účastníci zájezdu. Kód ve špičatých závorkách je identifikátor textu volně stažitelného v projektu Gutenberg (<www.gutenberg.org>).

V předchozím odstavci jsme užili naši experimentální metodu na vzorku textů různých jazyků segmentovaných v tomto případě na 2-gramy písmen. Náš experimentální postup však provádí stejnou analýzu za užití 1-gramů, 2-gramů, 3-gramů atd. (experimentálně jsme ověřili signifikantnost nejvýše 20-gramů písmen). Tento způsob analýzy nám umožňuje sledovat, jak postupně se zvětšující n-tice odrážejí strukturu textu, a to v určité kontinuitě proměn podoby grafu. Postupné zvětšování n-tic nám dává možnost sledovat strukturu textu na stále vyšších jazykových rovinách. Pozorujeme tak chování písmen (jakožto stanovené základní úrovně popisu), kombinací písmen, slov (daných jejich průměrnou délkou v určitém jazyce, pro češtinu je to 5-gram písmen; viz dále), dále vět atd. Všechny tyto jednotky jsou reprezentovány jako n-gramy písmen. Tento experimentální postup je plausibilní, v mnoha analýzách jsme ověřili, že průběhy grafů n-gramů odpovídajících průměrné délce dané jednotky v určitém jazyce (např. slov) jsou signifikantně podobné průběhům grafů těchto jazykových jednotek textu. Tuto proceduru nazýváme mapování.

Výsledkem procedury mapování určitého textu je *x* průběhů (zobrazujeme je do jednoho grafu), které můžeme použít pro porovnání s výsledkem procedury mapování jiného textu. Porovnáním se míní zjištění podobnosti jednotlivých navazujících n-gramových průběhů u obou textů. Například pokud známe jazykové jednotky ur-



OBRÁZEK 7: Mapování slovních forem českého textu. Text: M. Viewegh — Účastníci zájezdu.

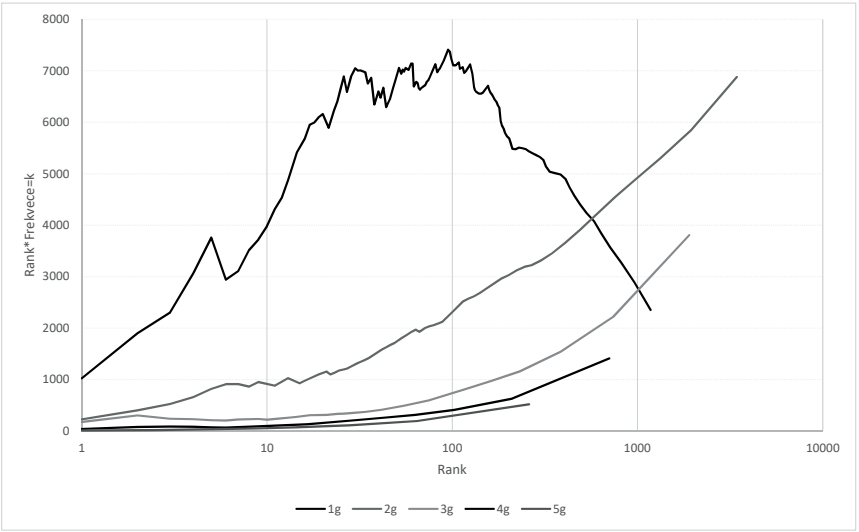


OBRÁZEK 8: Mapování písmen českého textu. Text: M. Viewegh — Účastníci zájezdu.

čitého textu, který dále zmapujeme, můžeme použít výsledný graf jako referenci k porovnání s výsledným grafem jiného textu, u kterého neznáme povahu jeho jazykových jednotek. Tímto způsobem se můžeme pokusit o jejich určení. Konkrétní příklady srovnání českého a čínského textu uvádíme níže.

První graf (obrázek 7) ukazuje mapování slovních forem českého textu. Druhý graf (obrázek 8) ukazuje mapování písmen rovněž českého textu. Na základě korelace jejich průběhů zjišťujeme, že průběh 1-gramů slov se podobá průběhu 4-gramů až 5-gramů písmen. Průběhy následujících n-gramů (u slov 2-gramy a vyšší, u písmen 5-gramy a vyšší) si svými průběhy také odpovídají. Korektnost korelace 1-gramů

slov s 4-gramy a 5-gramy písmen je zajištěna také podobností vývoje grafů v obou mapováních, nikoliv pouze podobností dvou konkrétních průběhů grafů. K nalezené hranici 4-gramů až 5-gramů písmen můžeme poznamenat, že průměrná délka českého slova, jak jsme experimentálně zjistili (na vzorku 243 českých literárních textů velikosti od 140 slovních forem do 149 070 slovních forem), má hodnotu 4,768 písmen. Srovnání obou uvedených mapování této hodnotě odpovídá, n-gramová analýza identifikuje hranici počtu písmen odpovídající slovům.

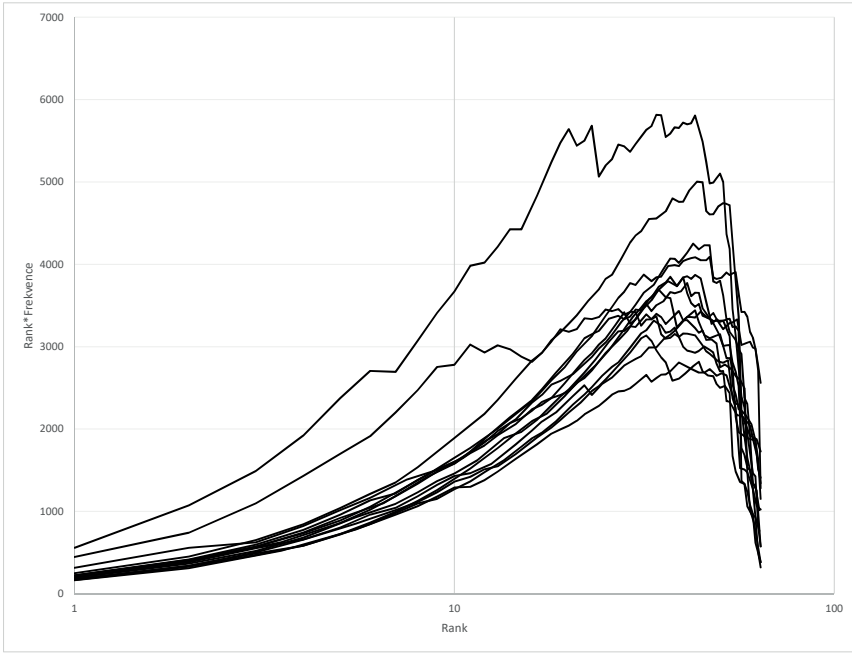


OBRÁZEK 9: Mapování znaků čínského textu. Text: čínština <25350-0>. Kód v ostrých závorkách je identifikátor textu volně stažitelného v projektu Gutenberg (<www.gutenberg.org>).

Na obrázku 9 je zobrazeno mapování znaků čínského textu. To srovnáme s předchozím mapováním písmen českého textu (obrázek 8). Sledujeme-li průběhy grafů mapování čínského textu, identifikujeme podobnost průběhu 1-gramů znaků čínského textu s 2-gramy až 3-gramy písmen českého textu. Podobně 2-gramy (a vyšší) čínského textu pak odpovídají průběhu grafů 5-gramů (a vyšších) písmen, respektive 2-gramů slov českého textu. Zjišťujeme tak, že znaky čínského textu nemají povahu písmen, ale spíš jejich dvoj- až trojkombinací, což odpovídá charakteru čínského znakového písma. Na základě výše stanovených kritérií porovnání průběhů grafů je korektnost tohoto závěru opět potvrzena.

OVĚŘENÍ METAFORY DNA

Výše uvedenou metodou mapování budeme analyzovat sekvence mRNA. Představili jsme standardní pojetí genetického kódu a různé analogie DNA a jazyka, včetně společných metod jejich zkoumání. Nejužívanější z analogií DNA a jazyka je pojetí bází DNA jakožto písmen — následně tripletů jakožto slov a genů jakožto vět (Jakobsonova

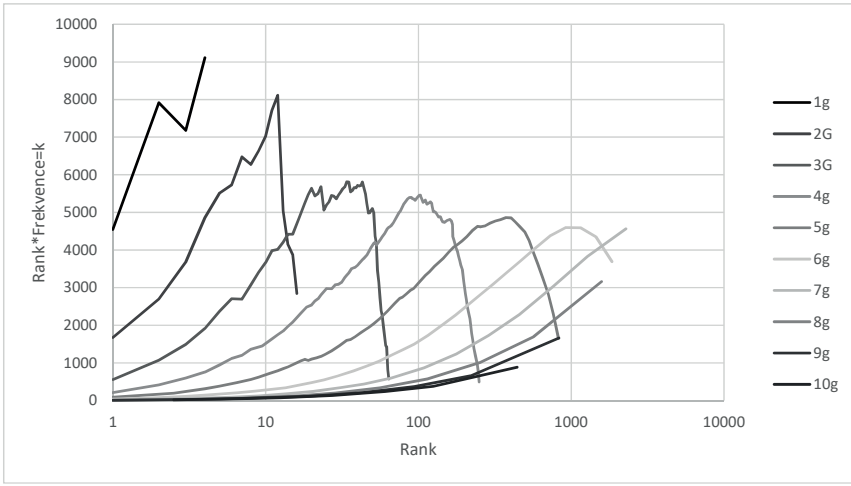


OBRÁZEK 10: Projev Zipfova zákona 3-gramů bází 20 náhodně vybraných sekvencí mRNA z analyzovaného vzorku. Text: ENA <AAA59187> 1 Homo sapiens (human) ras GTPase-activating-like protein, ENA <AAA59483> 1 Homo sapiens (human) epiligrin alpha 3 subunit, ENA <AAA59486> 1 Homo sapiens (human) laminin B1, ENA <AAA60554> 1 Homo sapiens (human) sodium channel alpha subunit, ENA <AAA59504> 1 Homo sapiens (human) lactase phlorizinhydrolase, ENA <AAA18895> 1 Homo sapiens (human) voltage-gated sodium channel, ENA <AAA51901> 1 Homo sapiens (human) calcium channel L-type alpha 1 subunit, ENA <AAA35629> 1 Homo sapiens (human) calcium channel alpha-1 subunit, ENA <AAA51898> 1 Homo sapiens (human) N-type calcium channel alpha-1 subunit, ENA <AAA15448> 1 Homo sapiens (human) DNA polymerase epsilon catalytic subunit, ENA <AAA60225> 1 Homo sapiens (human) protein tyrosine phosphatase zeta-polypeptide, ENA <AAA18639> 1 Homo sapiens (human) p300 protein, ENA <AAA59866> 1 Homo sapiens (human) mannose 6-phosphate receptor, ENA <AAA59924> 1 Homo sapiens (human) GAP-related protein, ENA <AAA58965> 1 Homo sapiens (human) collagen type VII, ENA <AAA52700> 1 Homo sapiens (human) heparan sulfate proteoglykan. Kód v ostrých závorkách je identifikátor textu volně stažitelného v genové bance EMBL-EBI (<www.ebi.ac.uk>).

metafora). Aplikací naší metody mapování vzorků sekvencí mRNA chceme ve srovnání s mapováním textů přirozeného jazyka prověřit, zda báze DNA a konsekventně mRNA hrají ve struktuře genetického kódu obdobnou roli jako písmena ve struktuře přirozených jazyků.

APLIKACE METODY

Výše představenou metodou bylo zmapováno přibližně 1000 náhodně vybraných mRNA sekvencí homo sapiens (EMBL-EBI, 2014; použité mRNA sekvence mají růz-

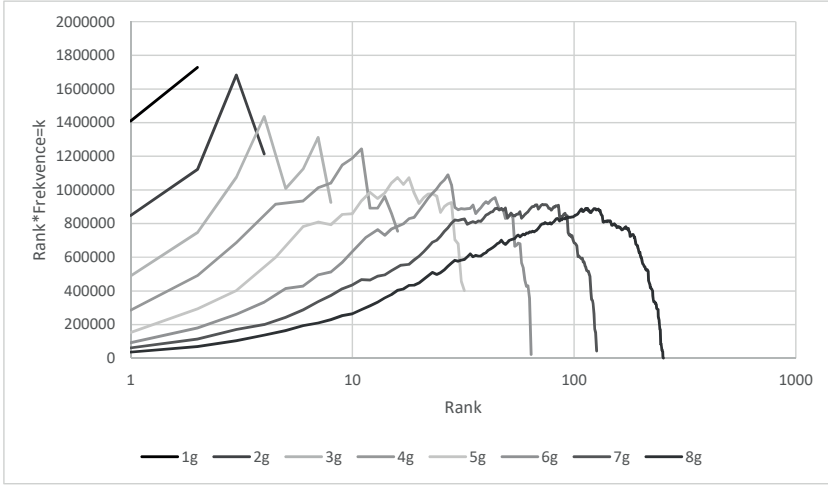


OBRÁZEK 11: mRNA. Text: ENA <AAA52700> AAA52700 1 Homo sapiens (human) heparan sulfate proteoglykan. Kód v ostrých závorkách je identifikátor textu volně stažitelného v genové bance EMBL-EBI (<www.ebi.ac.uk>).

nou délku, experimentálně jsme ověřili, že délka textu nemá vliv na mapování, pouze na obor hodnot, který není kritériem porovnání grafů). Zvolenou jednotkou je v této analýze báze. Aplikací naší metody (testováno na 1-gramech až 20-gramech) jsme zjistili signifikantní podobnost grafů různých sekvencí (viz např. obrázek 10 s grafy 3-gramů). Pro názornost grafu mapování však uvádíme analýzu pouze jedné náhodně vybrané sekvence (obrázek 11). Pro přehlednost ještě zopakujme, že obrázek 10 ukazuje 3-gramovou analýzu Zipfova zákona více vzorků mRNA a obrázek 11 1-gramovou až 10-gramovou analýzu Zipfova zákona jedné konkrétní sekvence.

Věnujme se nyní obrázku 11, který porovnáme s obrázkem 8 zobrazujícím mapování písmen českého textu. Můžeme si všimnout, že průběhy grafů mapování jsou si podobné od 3-gramů bází u DNA a 1-gramů písmen českého textu. Průběh 2-gramů bází je vůči průběhu 1-gramů písmen neúplný a až 3-gram bází realizuje křivku, kterou nacházíme u 1-gramů písmen (viz obrázek 12 zobrazující mapování distinktivních rysů hlásek, o kterých hovoříme níže — i zde je rozhodujícím kritériem úplnost průběhu grafu a podobnost průběhu grafu jako taková). Z tohoto důvodu klademe hranici písmene k 3-gramům bází. Následující průběhy obou mapování mají podobný vývoj. Průběhy 1-gramů a 2-gramů bází DNA nejsou u písmen realizovány, 3-gramy bází odpovídají 1-gramům písmen, 6-gramy až 7-gramy bází odpovídají 4-gramům až 5-gramům písmen. Další průběhy mají totožnou povahu.

Z průběhu obou mapování můžeme implikovat následující. Báze jsou jednotkou konstitučně nižší než písmena. U 2-gramů bází můžeme sledovat podobný, avšak neúplný průběh, jako mají 1-gramy hlásek (viz obrázek 12 zobrazující mapování distinktivních rysů hlásek, kde totéž platí pro 1-gramy až 5-gramy distinktivních rysů). To je vysvětlitelné degenerovaností genetického kódu, kde je pro kódování aminokyseliny často třetí báze redundantní (obdobně mnoho hlásek odlišují pouze jeden či dva distinktivní rysy). Trojicím bází (tripletům) přisuzujeme na základě podobnosti



OBRÁZEK 12: Mapování distinktivních rysů českého textu.

průběhů grafů roli písmen, sedmice bází (tedy více než 2 triplety) již tvoří obdobný typ konstituentu jako čtveřice až pětice písmen — tj. tvoří obdobu slov.

Vezměme první implikaci, která říká, že báze jsou jednotkou konstitučně nižší než písmena. Z této implikace vyplývá otázka, jakou roli hrají báze v genetickém kódu, analogizujeme-li jej s přirozeným jazykem a předpokládáme-li obdobný design obou kódů.

Písmena zcela intuitivně vnímáme jako nedělitelná. Tvořena jsou ovšem na základě vzájemných vztahů, které diferencují jedno písmeno od druhého. Každé písmeno je pak tvořeno souborem vlastností, který jej charakterizuje a zároveň odlišuje od ostatních. Lingvistika tyto vlastnosti nazývá distinktivními rysy. Písmena textu tak můžeme chápat jako soubory distinktivních rysů. S touto rovinou jsme při analýze přirozeného jazyka prozatím nepracovali.

Pro zavedení roviny distinktivních rysů do analýzy postupujeme následujícím způsobem: každé z písmen charakterizujeme jeho vlastnostmi (distinktivními rysy), které jej odlišují od ostatních. Neužili jsme tradičně lingvistikou popisované distinktivní rysy ve smyslu akustických vlastností hlásek, kterých je ve fonologických popisech více než deset. Použili jsme nejmenší možný počet k písmenům arbitrárně přiřazených distinktivních rysů schopných odlišovat písmena češtiny. Pro českou abecedu je takových opozic nutných pouze šest. V textu je každé písmeno reprezentováno unikátním řetězcem šesti pozic obsazených jedničkou nebo nulou (tedy přítomných nebo nepřítomných vlastností arbitrárních distinktivních rysů). Takto nově reprezentovaný text zmapujeme a porovnáme s výsledky mapování mRNA.

Výsledek mapování distinktivních rysů českého textu (viz obrázek 12) nám odhaluje typy průběhů, které se projevily u 1-gramů a 2-gramů bází DNA (obrázek 11). U mapování písmen podobné průběhy nenacházíme. Průběhům 1-gramů až 5-gramů distinktivních rysů písmen českého textu ovšem odpovídají průběhy grafů 1-gramů a 2-gramů bází. U distinktivních rysů je z důvodu jejich počtu mezi průběhy pozvolnější přechod. Průběhu grafu 6-gramů distinktivních rysů odpovídá průběh 3-gramů

bází. Z toho můžeme usuzovat, že báze mají v designu genetického kódu obdobnou roli jako distinktivní rysy písmen v designu přirozeného jazyka.

ZÁVĚR

Na základě aplikace Zipfova zákona na n-gramy bází DNA, písmena textu přirozeného jazyka a na distinktivní rysy písmen textu přirozeného jazyka můžeme usuzovat, že nukleotidové báze DNA plní v designu genetického kódu roli analogickou distinktivním rysům písmen textu přirozeného jazyka. Triplety bází jsou dále analogické písmenům. Kombinace tripletů jsou následně analogií slov.

Naším cílem bylo ověření analogie DNA a přirozeného jazyka. Jakobson formuloval analogii bází DNA a písmen, tripletů a slov, genů a vět. Naše analýzy tento design genetického kódu zpochybňují a ukazují analogii bází DNA s distinktivními rysy, tripletů s písmeny a kombinací tripletů se slovy. Všeobecně užívaná analogie bází jakožto písmen genetického textu se tak jeví jako chybná. Reprezentace genetického zápisu sledem písmen odpovídajících bázím (A, C, G, T) zřejmě zapříčinila pevné ukotvení této analogie ve vědecké praxi a v laickém pojímání genetického kódu. Naše výsledky však jasně ukazují, že tato analogie je vzhledem k designu přirozeného jazyka nesprávná a že „báze nejsou písmena".

LITERATURA:

Alberts, Bruce — Johnson, Alexander — Lewis, Julian — Raff, Martin — Keith, Roberts — Walter, Peter (2008): *Molecular Biology of the Cell* [5. vydání]. New York, NY: Garland Science.

Barbieri, Marcello (2002): *The Organic Codes: An Introduction to Semantic Biology*. Cambridge: Cambridge University Press.

Berg, Jeremy M. — Tymoczko, John L. — Stryer, Lubert (2012): *Biochemistry*. New York, NY: W. H. Freeman.

Bolshoy, Alexander — Volkovich, Zeev (Vladimir) — Kirzhner, Valery — Barzily, Zeev (2010): *Genome Clustering from Linguistic Models to Classification of Genetic Texts*. Berlin — Heidelberg: Springer.

Collado-Vides, Julio (1992): Grammatical model of the regulation of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 89(20), s. 9405–9409.

Collado-Vides, Julio (1993): A linguistic representation of the regulation of transcription initiation. I. An ordered array of complex symbols with distinctive features. *BioSystems*, 29(2–3), s. 87–104.

Crick, Francis H. C. (1962): Towards the genetic code. *Discovery*, 22, s. 8–16.

Crick, Francis H. C. (1964): On the genetic code: Nobel lecture, December 11, 1962. In: *Nobel Lectures: Physiology or Medicine: 1942–1962*. Singapore — New Jersey, NJ — London — Hongkong: World Scientific, s. 811–821.

Crick, Francis H. C. (1967): The Croonian Lecture, 1966: the genetic code. *Proceedings of the Royal Society of London B: Biological Sciences*, 167(1009), s. 331–347.

Crick, Francis H. C. (1968): The origin of the genetic code. *Journal of Molecular Biology*, 38, s. 367–379.

Cvrčková, Fatima (2006): *Úvod do praktické bioinformatiky*. Praha: Academia.

Faltýnek, Dan (2012): *Sémiotické primitivy v konstrukci gramatik: Testování gramatik jazyka a DNA*. Olomouc: Univerzita Palackého v Olomouci.

Ferrer-i-Cancho, Ramon (2006): When language breaks into pieces: a conflict between communication through isolated signals and language. *BioSystems*, 84(3), s. 242–253.

Ferrer-i-Cancho, Ramon — Elvevåg, Brita (2010): Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE* [online], 5(3). Cit. 8. 1. 2016. Dostupné z WWW: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009411>.

Ferrer-i-Cancho, Ramon — McCowan, Brenda (2009): A law of word meaning in dolphin whistle types. *Entropy* [online], 11(4), s. 688–701. Cit. 8. 1. 2016. Dostupné z WWW: <http://www.mdpi.com/1099-4300/11/4/688>.

Hartl, Daniel L. — Ruvolo, Maryellen (2013): *Genetics: Analysis of Genes and Genomes* [8. vydání]. Burlington, MA: Jones and Bartlett Learning.

Havlin, Shlomo — Buldyrev, Sergey V. — Goldberger, Ary L. — Mantegna, Rosario N. — Peng, Chung-Kang — Simons, Michael — Stanley, H. Eugene (1995): Statistical and linguistic features of DNA sequences. *Fractals*, 3(2), s. 269–284.

Jakobson, Roman (1971): Linguistics in relation to other sciences. In: Roman Jakobson, *Selected Writings: Vol. 2: Word and Language*. The Hague — Paris: Mouton, s. 655–696.

Ji, Sungchul (1997): Isomorphism between cell and human languages: molecular biological, bioinformatic and linguistic implications. *BioSystems*, 44(1), s. 17–39.

Ji, Sungchul (1999): The linguistics of DNA: words, sentences, grammar, phonetics, and semantics. *Annals of the New York Academy of Science*, 870, s. 411–417.

Ji, Sungchul (2006): The proteome as an autonomous molecular language: "proteinese"; A Poster accepted for presentation at the DIMACS Workshop on Sequences, Structure and Systems Approaches to Predict Protein Function, Rutgers University, Piscataway, NJ.

Katz, Gregory (2008): The hypothesis of a genetic protolanguage: an epistemological investigation. *Biosemiotics*, 1(1), s. 57–73.

Konopka, Andrzej K. (1995): Noncoding DNA, Zipf's law, and language. *Science*, 268(5212), s. 789.

Li, Wentian (1992): Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6), s. 1842–1845.

Mantegna, Rosario N. — Buldyrev, Sergey V. — Goldberger, Ary L. — Havlin, Shlomo — Peng, Chung-Kang — Simons, Michael — Stanley, H. Eugene (1994): Linguistic features of noncoding sequences. *Physical Review Letters*, 73(23), s. 3169–3172.

Mantegna, Rosario N. — Buldyrev, Sergey V. — Goldberger, Ary L. — Havlin, Shlomo — Peng, Chung-Kang — Simons, Michael — Stanley, H. Eugene (1995): Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Physical Review: E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 52(3), s. 2939–2950.

Markoš, Anton (1997): *Povstání živého tvaru*. Praha: Vesmír.

Markoš, Anton (2002): *Readers of the Book of Life: Contextualizing Developmental Evolutionary Biology*. New York, NY: Oxford University Press.

Markoš, Anton — Faltýnek, Dan (2011): Language metaphors of life. *Biosemiotics*, 4(2), s. 171–200.

Monod, Jacques (1970): *Le hasard et la nécessité: Essai sur la philosophie naturelle de la biologie moderne*. Paris: Éditions du Seuil.

Niyogi, Partha — Berwick, Robert C. (1995): A note on Zipf's law, natural languages, and noncoding DNA regions [online]. *A. I. Memo*, (1530) / *C.B.C.L. Paper*, (118). Cit. 8. 1. 2016. Dostupné z WWW: <ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-1530.pdf>.

Palazzo, Alexander F. — Gregory, T. Ryan (2014): The case for junk DNA. *PLoS Genetics*, 10(5).

Popov, O. S. — Segal, Daniel M. — Trifonov, Edward N. (1996): Linguistic complexity of protein sequences as compared to texts of human languages. *BioSystems*, 38(1), s. 65–74.

Raible, Wolfgang (2001): Linguistics and genetics: systematic parallels. In: Martin Haspelmath — Ekkehard König — Wulf Oesterreicher — Wolfgang Raible (eds.), *Language Typology and Language Universals: An International Handbook / Sprachtypologie und sprachliche Universalien: Ein internationales Handbuch / La typologie des langues et les universaux linguistiques: Manuel international*. Berlin — New York, NY: Walter De Gruyter, s. 103–123.

Searls, David B. (2002): The language of genes. *Nature*, 420, s. 211–217.

Stanford, Augustus L. (1975): *Foundations of Biophysics*. New York, NY: Academic Press.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, s. 57–74.

The european bioinformatics institute (EMBL-EBI) (2014). Cit. 19. 10. 2014. Dostupné z WWW: <http://www.ebi.ac.uk/ena/data/warehouse/search?query=tax_eq(9606)&domain=coding&result=coding_release&display=fasta&download=zip>.

Trifonov, Edward N. (1988): Codes of nucleotide sequences. *Mathematical Biosciences*, 90(1–2), s. 507–517.

Tsonis, Anastasios A. — Elsner, James B. — Panagiotis, A. Tsonis (1997): Is DNA a language? *Journal of Theoretical Biology*, 184(1), s. 25–29.

Twyman, Richard M. (1998): *Advanced Molecular Biology: A Concise Reference*. New York, NY — Abingdon: Taylor and Francis.

Twyman, Richard M. (2004): *Principles of Proteomics*. Abingdon — New York, NY: Garland Science / BIOS Scientific Publishers.

Watson, James D. — Berry, Andrew (2003): *DNA: The Secret of Life*. New York, NY: Alfred A. Knopf.

Weaver, Robert F. (2002): *Molecular Biology*. Boston, MA: McGraw-Hill.

Zipf, George Kingsley (1949): *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley Press.

Vladimír Matlach | Katedra obecné lingvistiky FF UP
<vladimir.matlach@upol.cz>

Dan Faltýnek | Katedra obecné lingvistiky FF UP
<dan.faltynek@upol.cz>