UNIVERZITA KARLOVA V PRAZE

Přírodovědecká fakulta

Katedra filosofie a dějin přírodních věd

Lingvistické přístupy v genomice a lingvistická metafora v biologii

Linguistic-like approaches in genomics and linguistic metaphor in biology

Disertační práce

Ing. Michaela Nohejlová Zemková

Vedoucí disertační práce: Prof. Jaroslav Flegr

Praha 2016

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze dne .....

.....

Michaela Nohejlová Zemková

Poděkování

Děkuji svému školiteli Jaroslavu Flegrovi za trpělivé vedení a všechny podněty a nápady, které mě přivedly k mnoha zajímavým tématům. Dále patří největší dík doc. Danielu Zahradníkovi z ČZU v Praze, bez jehož pomoci s matematickou a programátorskou stránkou práce by moje disertace stěží vznikla. Děkuji rovněž Prof. Edwardu N. Trifonovovi z university v Haifě za jeho velkorysé přijetí a dlouhodobou spolupráci, Fatimě Cvrčkové a Martinovi Mokrejšovi za bioinformatické konzultace, Ivanovi Čepičkovi za konzultace v oblasti zoologie. Dále děkuji Tondovi Markošovi, Zdeňkovi Neubauerovi a Zdeňkovi Kratochvílovi za jejich „myšlenkovou atmosféru“, díky které jsem se dostala za hranice běžné biologie (ne-li vždy v této práci, tak alespoň myšlenkově). Děkuji svým milým rodičům za nasměrování, které mi v životě dali, svému manželovi a dětem, že to se mnou během práce vydrželi.

Abstrakt:

V předložené práci jsou představeny články, jejichž spojujícím tématem jsou tzv. lingvistické přístupy v genomice, které umožňují genetické sekvence zkoumat jako „text“ obsahující potenciální „slova“ (oligonukleotidy, oligopeptidy) délky *n*. Tento přístup stojí na pomezí kvantitativní a kvalitativní analýzy a, na rozdíl od standardních komparativních metod užívaných v bioinformatice, umožňuje porovnání i fylogeneticky vzdálených jedinců. Stěžejním článkem práce (Zemková et al., 2016, submitted) je analýza peptidických slovníků parazitů a volně žijících organismů, která odhalila signifikantní rozdíly v diverzitě peptidů o délkách 4-6 aminokyselin u těchto porovnávaných skupin. Parazité obecně vykazují redukci pentapeptidů, která je částečně kompenzována zvýšenou diverzitou hexapeptidů. Tento výsledek je v souladu s námi postulovanou hypotézou, že parazité se snaží uniknout před imunitním systémem obratlovčího hostitele založeného na MHC imunitním rozpoznávání. Výsledek rovněž ukazuje, že klíčovou oblastí pro rozpoznání antigenu jsou peptidy o délce 4-5 aminokyselin a do reakce s T-receptorem tedy nevstupuje celý peptidový řetězec vázaný na MHC.

V dalších dvou článcích, které vzešly jako produkt spolupráce s Prof. E.N. Trifonovem z university v Haifě, je opět užita analýza slovníku. V prvním článku (Zemková et al., 2014) jde o detekci potenciálně amfipatických struktur, jejichž distribuce v proteomu vykazuje obecné zákonitosti, ale je zároveň pravděpodobně druhově specifická, takže by jí mělo jít použít jako tzv. „*proteomic signature*“ (Pietrokovski et al. 1990). Preference amfipatických peptidů a princip jejich formace je v článku dále analogizován se vznikem slov lidského jazyka, kde alternují (díky přirozeným omezením vyslovitelnosti určitých kombinací hlásek) souhlásky a samohlásky. Druhý článek (Zahradník et al., 2015) se zabývá potenciální možností rekonstrukce ancestrálních sekvencí DNA ze současného genomu. Je navržen algoritmus hledání tzv. generátorů (*generator sequence*), od kterých se odvíjela evoluce genomu.

Poslední předložený článek (Trifonov and Zemková, 2015) analogizuje vznik DNA z jednoduchých repetitivních sekvencí invazivního charakteru (Frenkel and Trifonov, 2012; Trifonov and Bettecken, 1997) a vznik lidské řeči z repetitivních holofrází tzv. kanonického žvatlání (*canonical babbling*). Extenzí této metafory vzniku života a vzniku jazyka/řeči je rovněž paralela v historii vědy: Obdobně jako došlo v přírodních vědách ke generalizaci genotypu jakožto jediné invariantní roviny vhodné pro výklad uspořádání živých bytostí (Např. Monod, 1971), byla i v lingvistice řeč jako individuální autonomní projev nahrazena

abstraktním systémem jazyka, jehož schopnost je dána vnitřním inherentním schématem (Chomsky 1986).

Literatura:

Chomsky N. 1986. Knowledge of language: Its nature, origin, and use. Greenwood Publishing Group.

Frenkel ZM, Trifonov EN. 2012. Origin and evolution of genes and genomes. Crucial role of triplet expansions. J Biomol Struct Dyn. 30(2):201-210.

Monod J. 1971. Chance and necessity. New York: Alfred A.Knopf.

Pietrokovski S, Hirshon J, Trifonov E. 1990. Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. Journal of Biomolecular Structure and Dynamics. 7(6):1251-1268.

Trifonov E, Bettecken T. 1997. Sequence fossils, triplet expansion, and reconstruction of earliest codons. Gene. 205(1):1-6.

Trifonov EN, Zemková M. (2015) Genome and language – two scripts of heredity (Ontogenetic theory of language origin). Czech and Slovak linguistic review

Zahradník D, Trifonov EN, Zemková M. 2015. Evolutionary landscape of human genome vocabulary. Czech and Slovak linguistic review

Zemkova M, Trifonov EN, Zahradnik D. 2014 One common structural feature of "words" in protein sequences and human texts. J Biomol Struct Dyn;32(7):1085-91.

Zemková M, Zahradník D, Mokrejš M, Flegr J. 2016. Reduction of peptide vocabulary in parasites – an evolutionary immune evasion strategy? (submitted)

Abstract:

In this work we present articles which are connected by the topic of linguistic-like approaches in genomics, which allow to treat genetic sequences as a "text" containing potential "words" (oligonukleotides, oligopeptides) of length *n.* Such an approach stands on the border of quantitative and qualitative analysis and, contrary to standard comparative bioinformatics methods, it is possible to compare phylogenetically distant individuals. Central article of my work (Zemková et al., 2016) is an analysis of peptide vocabularies of parasites and free-living organisms which showed significant differences in diversity of 4-6 amino acids long peptides of these compared groups. Parasites generally display reduction of pentapeptides, which is partly compensated by increased diversity of hexapeptides. This result is in accordance with our *a priori* hypothesis that parasites use immune evasion strategy to escape from MHC-based immunity system of its vertebrate host. Results also suggest that the length of key region for peptide recognition is about 4-5 amino acids and hence only short part of longer peptide bound in MHC participate on reaction with T-receptor.

In other two articles which arose as a product of cooperation with Prof Trifonov from the University of Haifa, we again used an analysis of genomic vocabularies. In the first article we detected potentially amphipathic structure. Distribution of these structures in proteome show some general regularities and they seem to be species – specific, so it could be possible to use them as so called "proteomic signatures" (Pietrokovski et al., 1990). Further we build an analogy between the preference of amphipathic peptides and the principle of its formation with the structure of words in human languages where consonant and vowels alternate due to natural constraints of pronounceability of their combinations.

Second article is dealing with possible reconstruction of ancestral sequences of DNA from recent human genome. An algorithm for potential reconstruction of the so-called generator sequence is suggested.

The last provided article (Trifonov and Zemková, 2015) builds an analogy between the origin of DNA from simple tandem repeats of invasive character (Frenkel and Trifonov 2012; Trifonov and Bettecken, 1997) and the origin of human speech from the utterances of so called *canonical babbling*. There is an extension of this metaphor of origin of life and language/ speech: In the history of science there is a parallel between natural sciences and linguistics. In both fields the living entities and languages were explained only from one invariant level of genome (Monod, 1971) or the language is considered to be an abstract system given by inner inherent scheme (Chomsky, 1986)

Obsah

Úvod

Základním zadáním, které jsem řešila během svého doktorského studia, byla otázka, zda se liší peptidické slovníky parazitů a organismů volně žijících, tedy otázka, kterou je možno zodpovědět prozkoumáním dostupných proteomů (celých sad proteinů)jednotlivých organismů. Zatímco standardní bioinformatické metody se zabývají posuzováním fylogenetické podobnosti a sdílením, respektive mírou podobnosti určitých částí genomu mezi organismy, těžištěm mojí práce byly tzv. „lingvistické genomické metody" (*linguistic-like approaches)*. Pomocí těchto metod lze zkoumat posloupnosti biologických makromolekul bez nutnosti předpokládané fylogenetické spřízněnosti a namísto vzájemného přiřazování sekvencí (*sequence allignement)* je celá posloupnost (genom, transkriptom, proteom) brána jako „text" v tom smyslu, že jej lze rozložit na potenciální slova délky *n (*oligonukleotidy/ oligopeptidy) a namísto s celou sekvencí pak pracovat s jejím „slovníkem" (*nukleotide/peptide vocabulary).* Tento přístup umožňuje jak kvantitativní analýzu – např. pomocí Shannonovy entropie a dalších odvozených indexů diverzity slovníků, tak kvalitativní analýzu, například míru užívání vzácných, nebo naopak široce sdílených „slov" u jednotlivých organismů. Všechny tyto metody (ať již vědomě či nevědomě) vycházejí z původního Shannonova konceptu (1948) rozkladu posloupnosti diskrétních prvků na tzv. *n*-gramy - slova/podřetězce o *n* znacích. Aplikace Shannonova konceptu je možná na jakoukoliv posloupnost diskrétních prvků – tedy i na genetické sekvence nebo na lidské jazyky. Skutečnost, že jak biologické makromolekuly nesoucí informaci, tak lidské zápisy jazyka, nejsou původně digitálním záznamem, ale entitou vyžadující tělesnost (jež jsou převodem na informační posloupnost vždy nějak redukovány), mě přivedla k extenzi tématu k obecným otázkám aplikace „lingvistických metafor v biologii" (Markoš et al., 2010). Tyto úvahy jsou shrnuty v příspěvku ke knize Biosémiotika II. (Markoš et al., 2015), jež vznikla na základě grantové spolupráce s UP v Olomouci v rámci projektu "Inovace studia obecné jazykovědy a teorie komunikace ve spolupráci s přírodními vědami".

Během doktorského studia jsem se věnovala především výzkumu peptidických slovníků jednotlivých organismů a problematikou jejich statistického porovnání. Výsledky této práce jsou shrnuty v publikaci Reduction of peptide vocabulary in parasites – an evolutionary immune evasion strategy? (Zemková et al., 2016) Vzhledem k tomu, že se jedná o mou klíčovou práci, která navíc vyžaduje širší osvětlení problematiky evoluce parasitismu a imunitního systému, včetně matematického pozadí řešení problému, je jí v mé disertační práci věnováno nejvíce prostoru. Dále jsou přiloženy dvě publikace ve spoluautorství s Prof. E. N. Trifonovem z University v Haifě a Doc. Danielem Zahradníkem z ČZU v Praze: One

common structural feature of "words" in protein sequences and human texts (Zemková et al., 2014) a Evolutionary landscape of human genome vocabulary (Zahradník et al., 2015), kde je využito *n*-gramové analýzy. V prvním článku je publikován software pro detekci potenciálně ampfipatických struktur v proteinech a počítání s *n*-gramy libovolné délky. Článek je následně postaven na hříčce mezi strukturní analogií amfipatických struktur, kde se ve specifické periodě střídají polární a nepolární aminokyseliny, podobně jako se střídají souhlásky a samohlásky v projevech lidských jazyků. Je třeba zdůraznit, že tato hříčka si nebere za cíl porovnávat jakkoliv kvantitativně genetické a lidské "texty". (Nehledě na to, že by to nemělo smysl filosofický, jsou zde metodické překážky při porovnání textů v cizích jazycích, vzniklé nemožností jednotné fonetické transkripce.)Jde pouze o ilustraci skutečnosti, že (obě tyto) složité informační struktury vznikají zavedením základního fyzického omezení pocházejícího z tělesného světa: tedy polaritou (respektive hydrofobií) jednotlivých aminokyselin v případě struktur amfipatického helixu a místa a způsobu tvorby hlásek v případě lidských slov. Přestože „podobnost" struktur je zde spíše artefaktem digitalizace jejich původní tělesné podoby, je třeba dodat, že i v reálném fyzickém světě pak pozorujeme existenci těchto struktur jakožto znaků – slov nebo prostorových struktur nesoucích informaci.

Druhý článek se věnuje potenciální možnosti rekonstrukce evoluční historie genomu z jednoduchých tandemových repetic, kterou navrhuje Frenkel a Trifonov (2012). Přestože množství genomických dat produkovaných moderními skenovacími technologiemi je obrovské, je překvapivě málo studií, které by tato data využily k odhalování evoluční historie genomu. Základní myšlenka je velmi obecná-formulovaná v evoluční biologii např. Susumu Ohnem v jeho známém díle Evoluce genovou duplikací (1970) - vznik komplexních struktur z jednoduchých opakujících se nukleotidových motivů. V článku se pokoušíme na příkladu lidského genomu ilustrovat, jak by bylo možné pomocí *n*-gramové metody rekonstruovat prastaré jednoduché sekvence (tzv. *sequence generators*), od kterých se odvíjela následná evoluce genomu.

Kromě těchto dvou článků, které jsou založeny na *n*-gramovém počtu, je myšlenka evoluce komplexních struktur z jednoduchých opakujících se motivů rozvinuta v článku Genome and language – two scripts of heredity (Ontogenetic theory of language origin) (Trifonov and Zemková, 2015). Jde sice opět o „hříčku" analogizování evoluce genomu z jednoduchých repetic a povstávání lidského jazyka z dětského repetitivního žvatlání (tzv. *canonical babbling*), ale při hlubším nahlédnutí do lingvistického zákulisí se ukazuje, že mechanistické chápání evoluce zabraňuje lingvistům optimálně sloučit jazyk a řeč (tedy abstraktní reprezentaci a individuální tělesný projev) podobně jako digitální genetický zápis a živou

bytost v biologických vědách. Díky tomu bylo téma evoluce jazyka již koncem 19. století oficiálně odsunuto mimo akademickou půdu a znovu se dostává do popředí až s rozvojem kognitivních přístupů koncem 80. let (Fitch, 2005), ale i v současnosti je hlavním proudem lingvistiky (reprezentovaným např. N. Chomskym) označováno jako kontroverzní a zkoumání dětského jazyka není považováno za přínosné pro otázku vzniku jazyka obecně. Analogie zde tedy má přesah i do historie vědy. Článek vznikl na základě spolupráce s katedrou obecné lingvistiky UP v Olomouci a jeho základní myšlenky byly prezentovány na symposiu „Informational fundamentals of life: genome and languages“ v Olomouci 2014.
Ostatní práce, které se však tematicky nijak nedotýkají výše uvedených témat, jsou uvedeny v seznamu ostatních publikací.

Literatura:

Frenkel ZM, Trifonov EN. 2012. Origin and evolution of genes and genomes. Crucial role of triplet expansions. J Biomol Struct Dyn. 30(2):201-210.

Fitch WT. 2005. The evolution of language: a comparative review. Biology and philosophy.20 (2-3): 193-203

Markoš A, eds. 2010. *Jazyková metafora živého*. Červený Kostelec: Pavel Mervart.

Markoš A, Kleisner K, Stella M, Zemková M. 2014. Biosémiotika II. UP v Olomouci

Ohno S. 1970. Evolution by gene duplication. Springer Science & Business Media.

Shannon CE., 1948, A mathematical theory of communication: The Bell System Technical Journal, v. 27, p. 379–423, 623–656.

Trifonov EN, Zemková M. 2015. Genome and language – two scripts of heredity (Ontogenetic theory of language origin). Czech and Slovak linguistic review.

Zahradník D, Trifonov EN, Zemková M. 2015. Evolutionary landscape of human genome vocabulary. Czech and Slovak linguistic review

Zemkova M, Trifonov EN, Zahradnik D. 2014 One common structural feature of "words" in protein sequences and human texts. J Biomol Struct Dyn;32(7):1085-91.

Zemková M, Zahradník D, Mokrejš M, Flegr J. 2016. Reduction of peptide vocabulary in parasites – an evolutionary immune evasion strategy? (submitted)

Reduction of peptide vocabulary in parasites – an evolutionary immune evasion strategy?

1. Úvod

Parazit je rozpoznáván v těle hostitele podle přítomnosti cizorodých peptidů, které se váží u obratlovců s vývojově pokročilejší imunitou do komplexu tzv. MHC molekul na povrchu většiny somatických buněk. Dalo by se tedy předpokládat, že parazité budou mít tendenci redukovat rozmanitost svých peptidických „slov“ a zbavit se všech postradatelných peptidů, aby lépe unikli před imunitním systémem hostitele. Tato hypotéza je zde testována na základě analýzy peptidických slovníků 30 jednobuněčných endoparazitů, 17 volně žijících jednobuněčných organismů, 8 mnohobuněčných parazitů a 16 volně žijících mnohobuněčných. Porovnáním těchto slovníků v rozmezí délek slov 4-12 aminokyselin jsme zjistili, že paraziti (jednobuněční i mnohobuněční) mají významně ochuzenou diverzitu pentapeptidů a naopak zvýšenou diverzitu hexapeptidů než volně žijící organismy. Jako potenciální činitelé ovlivňující bohatost slovníku byly vzaty v potaz i další faktory jako délka proteomu, počet proteinů, redundance atd., ale parazitismus se ukázal jako nejsilnější ze všech studovaných faktorů. Tyto výsledky podporují naší hypotézu, že redukce peptidického slovníku může být významnou únikovou strategií parazita před imunitním systémem hostitele. Rovněž naznačují, že pro vlastní imunitní rozpoznávání jsou kritické pouze délky peptidů 4-5 aminokyselin, namísto obvykle předpokládaných 7 – 12.

1.2. Rozpoznávání cizorodých peptidů v buňce

U organismů s vyspělým imunitním systémem jako jsou vertebrata, jsou vlastní a cizorodé proteiny rozpoznávány tak, že protein je uvnitř buňky degradován na krátké fragmenty a tyto peptidy jsou posléze vystaveny na povrchu buňky vázány do komplexu tzv. MHC molekul a vystaveny buňkám imunitního systému, které buňku s cizorodými peptidy buď přímo zlikvidují, nebo vyvolají imunitní odpověď hostitelského organismu. Pokud buňka není napadená žádným antigenem, T- receptor s prezentovanými peptidy nereaguje. (T-lymfocyty, které by reagovaly s vlastními peptidy, jsou odstraněny negativní selekcí během svého zrání v brzlíku, kde jsou testovány na autoreaktivitu.)

Je zajímavé, že zatímco receptory B-lymfocytů vážící se přímo na antigen, rozeznávají cizorodé agens na základě terciální struktury, T-receptory potřebují pro rozpoznání peptid v jeho denaturované – v podstatě lineární podobě (Berzofsky, 1989). Skutečnost, že peptidy jsou vybírány v buňce náhodně (Neefjes and Ovaa, 2013) (Vyas et al, 2008) a poté stříhány a denaturovány, nás opravňuje k tomu, abychom v našem experimentu považovali peptidy za lineární sled aminokyselin a zanedbali jejich složitou terciální strukturu. Rovněž vzorky peptidů, které jsou analyzovány, jsou náhodně – neselektivně- vybrány z celého proteomu.

Parazit je tedy v hostitelském organismu rozpoznáván pomocí imunitního systému podle specifických peptidů, které mají motiv odlišný od hostitelových vlastních peptidů. Použijeme-li lingvistickou analogii, můžeme říct, že peptidický „slovník" parazita a hostitele se liší. Modifikace peptidického slovníku tak může být významnou evoluční strategií úniku parazita před imunitním systémem hostitele.

Téměř všechny buňky v tělech obratlovců (kromě spermií a trofoblastů) mají na svém povrchu MHC molekuly a to v hustotě přibližně 10^5 -10^6 molekul. Somatické buňky disponují komplexem MHC I (MHC class I) zatímco specializované buňky imunitního systému tzv. buňky prezentující antigen (*Antigen presenting cells*, APC) jako dendritické buňky (DCs), makrofágy nebo B-lymfocyty mají na svém povrchu MHC II (MHC class II). Kromě patogenních jsou uvnitř buňky zpracovány i vlastní peptidy jako alternativní produkty translace, různé poškozené translační produkty (označované souhrnně jako *defective ribosomal products*), ale i proteiny pohlcené buňkou z vnějšího prostředí (Trombetta and Mellman, 2005). Ukazuje se rovněž, že dendritické buňky jsou schopné tzv. cross-presentace – tedy že degradují extracelulární agens a poté jej přenesou dovnitř buňky, kde podstoupí standardní cestu končící prezentací na MHC I, což je poměrně neobvyklé a kompletní cesta antigenu během tohoto procesu není úplně známa (Paz et al. 1999; Vyas et al. 2008). Z tohoto je zřejmé, že neplatí jednoduchý učebnicový předpoklad, že na MHC I jsou vystavovány pouze peptidy z vnitřního prostředí buňky.

1.2.1. Zpracování a prezentace antigenu na MHC I

Samotná molekula MHC I se skládá ze tří modulů: části upevňující celý komplex v buněčné membráně, imunoglobulinového modulu a modulu vázajícího peptid. Pro naše další úvahy je důležité především vazebné místo pro peptid (*peptide binding region*). To je tvořeno dvěma α-řetězci, které tvoří žlábek (*peptide binding groove*), jehož „dno" drží

struktura β-sheet. Peptid se váže dovnitř tohoto žlábku, takže část, která se dostane do kontaktu s T-receptory buněk imunitního systému je pouze tato omezená část „trčící“ ven, zatímco koncové části peptidového řetězce slouží k ukotvení peptidu ve struktuře MHC.

Peptidy pocházející z vnitřního prostředí jsou zpracovány v ER, kde jsou střihány na délky 8-10 aminokyselin, jsou vázány do komplexu nascentní molekuly MHC I a β2-microglobulinu (Vyas et al., 2008) (Neefjes and Ovaa, 2013). Toto spojení umožňuje optimální sbalení a transport celého komplexu na povrch buňky. Experimentálně bylo ukázáno, (Chang et al., 2005) že aminopeptidázy v ER ukončí střih (*trimming*) peptidů, když délka řetězce je menší než 8 aminokyselin.

Horní hranice délky peptidického řetězce je tedy poměrně jasně stanovena, nicméně na samotné interakci s T-receptorem se pak podílí pouze jakási jádrová část (přibližně uprostřed) denaturované molekuly peptidu. Kritická oblast se zdá být pouze 2-3 aminokyseliny dlouhá, ale vlastní odpověď T-receptoru je možná teprve připojením další 1-2 aminokyselin (Austyn and Wood 1994). Tedy vlastního imunitního rozpoznávání se na MHC I účastní jádrová část peptidu o průměrné délce asi 5 aminokyselin, zatímco koncové části slouží pouze jako vazebné.

1.2.2. Prezentace antigenu na MHC II

Struktura molekuly MHC II je obdobná jako u MHC I, obojí jsou syntetizovány v ER pouze na vazebném místě pro navázání peptidu je vázán tzv. invariantní řetězec, který blokuje vazebné místo, aby se na něj nenavázaly vlastní peptidy před tím, než dojde k vazbě s antigenním peptidem. Peptidy z vnějšího prostředí buňky jsou totiž zpracovány poněkud odlišnou cestou přes fagosom, který fúzí s lysozomem dává vzniknout fagolysozomu, ve kterém dochází k interakci s MHC II a poté jsou tyto exogenní peptidy prezentovány na MHC II na povrchu buňky. Ukazuje se, že kromě této „tradiční“ cesty, kdy je antigen nejdřív zpracován a teprve potom zabudován do MHC II je možný i scénář, kdy je peptid zachycen a vázán rovnou do vazebného žlábku a jeho přečnívající konce jsou poté odstřiženy lysozomálními proteázami (Trombetta and Mellman, 2005). Na rozdíl od MHC I je totiž vazebné místo pro peptid otevřeno na obou koncích a mohou se tak vázat i delší řetězce přímo na povrchu APC (Rammensee et al., 1995; Vogt and Kropshofer, 2011). Důležitá je skutečnost, že invariantní řetězec je postupně odštěpován, až z něj zůstane pouze malý fragment označovaný jako CLIP (Class II Associated Invariant chain Peptide). I tento

fragment má dvě části, obdobně jako antigen má svou vazebnou část (CLIP core region 15 aminokyselin dlouhý) a část prezentovanou T-receptoru (CLIP effector site – 9 aminokyselin dlouhý). (Vogt and Kropshofer, 2011)

1. 2. 3.Polymorfismus genů pro MHC

Geny pro MHC glykoproteiny jsou velmi polymorfní – mají velké množství alel, jež se liší hlavně v místě, které zodpovídá za vazebné místo pro antigen. Předpokládá se, že kolik variant pro vazebná místa existuje, tolik specifických antigenů je možno vázat. Polymorfismus MHC tedy zajišťuje možnost rozeznání co nejširšího spektra antigenů. Samotný objev MHC souvisel s klinickou praxí odhojování transplantátů – tedy že pacientův imunitní systém nepřijal buňky vystavující odlišný repertoár MHC glykoproteinů.

Přestože vlastní evoluce MHC se počítá od čelistnatých obratlovců výše, ortology MHC genů lze nalézt u evolučně nižších skupin včetně rostlin.(Kelley et al., 2005), (Kasahara et al., 2004) Podle Kelleyho se tyto ortology objevují již zhruba před 600 milióny let u společného předka rostlin živočichů a hub. Vlastní, již plně strukturované MHC, se objevuje až u paryb (cca 500 miliónů let). Vlastní oblast pro MHC geny má suboblasti I.- III. (class I – pro MHC I, class II pro MHC II a class III pro spolupůsobící geny včetně takových, které se na vlastním imunitním rozpoznávání nepodílejí, ale hrají úlohu v aktivaci zánětlivé reakce) (Kelley et al., 2003).

Komparativní genomické studie ukazují na velmi rychlou evoluci této oblasti genomu hned po oddělení se od společného předka obratlovců, neboť i velmi blízké druhy mají často diametrálně odlišnou organizaci i obsah jednotlivých oblastí. Lidský komplex pro MHC obsahuje přes 260 genů (Beck and Trowsdale, 2000). Šimpanzi a gorily jako nejbližší příbuzní mají velmi podobnou architekturu lidskému MHC a to zejména v oblasti I. Naopak nám vzdálenější primáti jako makak prodělali v této oblasti duplikaci. MHC myší sdílí lidskou architekturu, zatímco u krys se architektura v jednotlivých oblastech značně liší. Obecně se liší mezidruhově i vnitrodruhově počet genů pro MHC, vzájemná poloha oblastí a jejich obsah. Během velmi dynamické evoluce této oblasti docházelo hlavně k duplikacím a delecím. Některé moduly se vyvíjely společně, jako například transportní molekuly a jejich příslušné MHC. Oblast pro MHC I patří k nejvíc polymorfním a mezi jednotlivými druhy i jejich skupinami je velmi rozrůzněná (Kelley et al., 2003). Co však patří ke konzervovaným

částem MHC jsou geny pro strukturu a funkci molekul presentujících peptidy (Madden, 1995) (Kelley et al., 2003).

1.2.4. Evoluce adaptivní imunity

Existence nespecifické imunity zajišťující eliminaci cizorodých agens je známa jak u eukaryot, tak u prokaryot. Tento typ imunity založený na rozpoznání specifických cizorodých motivů a následném fagocytování a odstranění jejich původce patří k vývojově nejstarším typům imunity (Boehm and Swan, 2014)(Hořejší et al. 2013). Analýza receptorů pro mikrobiální struktury (*Pattern recognition receptors*), které mají rostliny, nižší i vyšší živočichové ukazuje na pravděpodobný společný základ těchto mechanismů u eukaryot. Jak živočichové, tak rostliny používají struktury podobné Toll-like receptoru a LRR receptoru, které jsou schopné vázat široké spektrum molekul a hrají roli i v evoluci adaptivní imunity jako stavební kameny imunitního systému bezčelistnatých. Obratlovci již za stejným účelem mají vyvinuté imunoglobuliny (Ronald and Beutler, 2010).

Evoluce adaptivní – specifické - imunity umožňující buněčnou „paměť“, díky které se při následném setkání s patogenem již odpověď organismu řídí předešlou zkušeností z infekce, je sice obvykle rekonstruována až od společného předka obratlovců, ale v určité formě je přítomna již u prokaryot. Přítomnost získané nebo též „trénované“ imunity je známa i od jiných skupin, než jsou obratlovci. U bakterií a archeí je znám mechanismus ochrany před cizí DNA založený na tzv. CRISP sekvencích (krátké palindromické sekvence), do kterých je zabudována cizí DNA a pokud se tato znovu objeví, je identifikována a zlikvidována. Tento systém má mimo funkce obranné zřejmě i funkci při udržování symbiotického vztahu s eukaryotickými hostiteli (Boehm and Swan, 2014).

Některé mechanismy tradičně řazené k nespecifické imunitě ale stojí na pomezí specifické imunity. Instruktivním příkladem jsou NK buňky („natural killers“), které sice rozpoznávají antigeny nespecificky, ale při setkání s antigenem se pomnoží a změní se v tzv. aktivovanou formu, která má potenciál paměťové buňky. Produkcí cytokinů a chemokinů zároveň ovlivňují procesy specifické imunity a obdobně jsou i ovlivňovány buňkami specifické imunity jako jsou lymfocyty (Kopecký and Kopecký, 2010) (Boehm and Swann, 2014). Evoluční propast mezi specifickou (evolučně mladší) a nespecifickou imunitou tedy zřejmě nebude až tak zásadní, přinejmenším je zřejmé, že v obou se uplatňují obecné principy pro rozeznávání cizího od vlastního, které se odrážejí i v používání osvědčených stavebních

kamenů jako jsou konzervované struktury některých receptorů. Prekursory NK buněk mají už sumky. Rod *Botryllus* zřejmě používá NK buňky obecně k rozpoznávání cizorodého. U urochordát (konkrétně u *Ciony intestinalis*) byl pozorován alternativní způsob rozpoznávání cizorodého a to nikoliv pomocí nějakého proto-MHC, ale pomocí struktur odpovídajících obratlovčím receptorům pro komplement. Konkrétně jde o receptory CD55 a CD46, které jsou exprimovány u obratlovců na buňkách v okolí, kde je komplement aktivován. (Aktivovaný komplement totiž může při zásahu v místě infekce napadnout i okolní „nevinné" buňky, tyto pokud mají výše zmíněné receptory, jsou rozeznány jako vlastní a proti komplementu jsou ochráněny.) Geny pro tyto receptory u *Ciony* vykazují neobvyklou vnitrodruhovou variabilitu, která není u obratlovců známa. Autoři tohoto objevu (Kürn et al., 2007) se domnívají, že tyto geny zajišťují rozpoznávání vlastního a cizího v absenci MHC a uplatňují se i při zabraňování samooplození či mezidruhovém křížení.

U mihule jsou známy homology podobné savčím lymfocytům. Už u bezčelistnatých nalézáme proto- formu lymfatických tkání jako thymus (označovaných někdy jako thymoid) (Boehm and Swan, 2014). Plně funkční imunitní systém zahrnující MHC rozpoznávání cizorodých peptidů mají čelistnatci, zatímco bezčelistnatí mají pouze ortology MHC. Tato skutečnost ukazuje, že evoluce adaptivní imunity se odehrála velmi bouřlivě. Zdá se, že z hlediska evoluce genomu byly pro rozvoj adaptivní imunity důležité dvě události: osvojení genu aktivujícího rekombinaci (*RAG*), jehož možný původ je odvozen od prokaryotického transpozonu, a který není přítomen u bezčelistnatých, a dvě vlny genomových duplikací stojících u zrodu obratlovců (Kasahara et al. 1997). Podle teorie první duplikace stála za oddělením vertebrat od společného předka a druhá za oddělením čelistnatců (Wolfe, 2001) (Ohno, 1968). Tato druhá vlna polyploidizace by mohla být zodpovědná za rychlý rozvoj MHC oblasti. Gen *RAG* obecně umožnil vznik variabilních oblastí, které jsou v pozadí širokého repertoáru buněk adaptivní imunity. (Kasahara et al. 2004)

1.3. Lingvistická analýza genetických sekvencí

Standardním bioinformatickým přístupem ke genetickým sekvencím je přiřazení (*alignment*) dvou nebo více sekvencí a hledání jejich vzájemné podobnosti a následné úvahy o jejich evolučních vztazích. Tento přístup lze ovšem uplatnit pouze u sekvencí, kde nějakou evoluční spřízněnost předpokládáme a výsledek vždy závisí na kvalitě zpracovávaného

přiřazení. (Bolshoy, 2003) Lingvistické metody (v angličtině označované častěji jako *linguistic-like tools*) oproti tomu umožňují porovnávat v podstatě libovolné sekvence. Bolshoy, který je patrně jediným teoretikem v této oblasti, je doslova označuje jako„*alignement-free* metody", jež se nezabývají přiřazením jednotlivých sekvencí, ale berou sekvenci jako text, který je možno rozložit na potenciální slova délky *n* (tzv. *n*-gramů), následně zkoumat jejich frekvence a distribuci ať už v rámci jednoho nebo více organismů, jednotlivých genů/proteinů, či celých genomů/proteomů.
Termín „lingvistický" tedy znamená spíše „založený na *n*-gramové analýze", která se v praxi používá v oblastech, jenž se obecně označují jako *NLP – Natural language processing* – což jsou např. převádění řeči na text (*speech recognition*), mechanické překlady, automatické opravy dokumentů či syntaktická analýza a další. Stejným způsobem jako můžeme takto rozložit text v přirozeném jazyce (ať už budeme brát v potaz existenci mezer jakožto znaků či nikoliv), lze rozložit i sekvence DNA, RNA či proteinů a zkoumat distribuci jednotlivých *n*-gramů. Analýzy ukazují, že i organismy mají kompozici a užívání svého „slovníku" nenáhodné a že existují „slova" typická pro dané skupiny organismů. Podobně jako u lidských textů hovoříme např. o autorském stylu, a tak se zavedl pro tyto typické vzorce termín *genomic signature.* Již na úrovni di-gramů – tedy rozkladu sekvence proteinu na slova (peptidy) délky 2 – se ukazuje, že lze vzájemně rozlišovat jednotlivé organismy. Dle průkopníků této metody (Pietrokovski et al., 1990) (Brendel et al., 1986) stačí ke spolehlivému určení nějakého organismu velikost *n*-gramu pro hodnoty $2 \leq n \leq 6$.

Lingvistické metody nejsou schopny plně nahradit standardní srovnávací bioinformatické metody založené na přiřazování a hledání homologií, ale mohou posloužit jako alternativní doplňující nástroje k detailnější analýze. Navíc jsou výpočetně méně náročné. Jak ukázal Karlin (1994) a bylo zkoumáno již v 80. letech, pro relevantní rozlišení dvou porovnávaných sekvencí DNA stačí porovnávat slovník tvořený slovy délky 4 (tedy tetraoligonukleotidy).

1.3.1. Indexy lingvistické komplexity a jejich využití[1]

Na sekvence biologických makromolekul jako jsou proteiny a DNA lze tedy pohlížet z hlediska klasické Shannonovy teorie informace (Shannon 1948) jako na lineární

[1] Poznamenejme, že termínem **komplexita** je zde vždy míněná diverzifikovanost – bohatost slovníku, nikoliv termín užívaný pro velikost genomu označovanou v literatuře jako C-value. (Flegr, 2003, str. 141)

posloupnost prvků schopných uchovávat informaci, jejíž potenciální obsah je možno vyjádřit pomocí matematických charakteristik, kde vstupními parametry jsou délka sekvence (*N*), rozsah abecedy (*s*) a délka podřetězců (*n*) – tzv. *n*-gramů („slov“ zvolené délky *n*, jejichž frekvenci chceme zkoumat). Počet všech možných *n*-gramů (oligonukleotidů, oligopeptidů) je s^n. Způsob, jakým jsou distribuovány konkrétní n-gramy a s jakou frekvencí, je možné vyjádřit pomocí různých matematických charakteristik. Nejjednodušší takovou charakteristikou je *Shannonova entropie*, která vyjadřuje informační potenciál sekvence:

$$H' = -\sum_{i=1}^{s} \frac{n_i}{N} \ln_s \frac{n_i}{N}$$

Což je vzorec izomorfní se vzorcem entropie, jak byla definována ve statistické mechanice Boltzmanem,[2] ale v našem případě jde o vyjádření výskytu i-tého znaku n_i na posloupnosti délky *N* a rozsahu abecedy *s*. Nejvyšší entropii má tedy sekvence, kde se vyskytují všechna písmena zcela náhodně. Nulovou entropii (nejvyšší uspořádanost, ale také nejnižší pravděpodobnost) by měla sekvence, kde se opakuje pouze jedno písmeno. Pro úvahy o genetických sekvencích pak má praktické aplikace tzv. *relativní entropie* neboli *redundance* (podíl entropie zdroje ku maximální entropii), která odráží míru repetitivnosti sekvence.

Informace se ale týká pouze frekvence písmen, nikoliv jejich distribuce uvnitř „textu“. [3]Tyto jemnější charakteristiky jsou schopny zachytit různé indexy, které obvykle nějak poměřují aktuální využití slovníku vůči všem možným existujícím kombinacím, které se obecně označují jako **indexy komplexity** (Více Orlov and Potapov, (2004); Zemková and Zahradník, 2015). Tyto indexy mají v praxi obecně tu výhodu, že jejich výstupem jsou čísla, která lze mezi sebou porovnat, a algoritmy jsou obvykle výpočetně nenáročné. Problematické jsou ale jejich výsledky na příliš dlouhých nebo naopak krátkých sekvencích a jejich různá citlivost na repetitivní úseky.

Příkladem praktického využití *n*-gramové analýzy je např. bioinformatický program BLAST, kde jsou implementovány algoritmy pro porovnávání slovníků, což výrazně urychluje prohledávání databází (Cvrčková, 2006). V tomtéž programu se uplatňuje také Wooton-Federhen index, který slouží k odhalování oblastí genomu s nízkou komplexitou (*LCR-Low complexity regions*)(Sharon et al. 2005).

---

[2] kde *pi* je definováno jako pravděpodobnost, že se systém nachází v i-tém stavu -i-té buňce- svého fázového prostoru a vzorec navíc obsahuje univerzální Boltzmannovu konstantu *K* .

[3] Například sekvence AAAACCCCCC bude mít stejnou entropii jako ACACACCACA: $H' = 4/10 * \log_2(4/10) + 6/10 * \log_2(6/10)$ (N=10, n_1 =4, n_2 =6. Abeceda je dvoupísmenná, A se vyskytuje 4krát, C 6krát).

Změny komplexity slovníku lze sledovat přímo na sekvenci tak, že si na posloupnost přiložíme „okénko“ *(sliding Windows)* o zvolené délce (obvykle 20 písmen), a postupně ho posouváme po sekvenci. Pro každé okénko pak zaznamenáváme hodnotu komplexity. Takto lze například ukázat, že se významně mění lingvistická komplexita na rozhraní intronů a exonů, což je možno následně využít v predikčních algoritmech. (Bolshoy 2003; Orlov and Potapov 2004)

1.3.2. Využití slovníku a jeho obecné vlastnosti

Zatímco indexy lingvistické indexy poskytují komprimovanou informaci o celkové komplexitě pro všechna slova ze zvoleného rozmezí *n*, pro naše analýzy bylo třeba se zabývat využitím slovníku konkrétní délky. Zatímco srovnání celkové komplexity pomocí Shannonovy entropie a Trifonovova indexu (Popov et al. 1996) neodhalilo žádné rozdíly v lingvistické komplexitě námi studovaných organismů, na úrovni konkrétních délek peptidů se již významné rozdíly vyjevují. To se ukazuje jako smysluplné vzhledem k mechanismu MHC- imunitního rozpoznávání, jež je založeno na prezentaci peptidů T-receptorům. Tyto peptidy jsou vystaveny na komplexu MHC molekul v denaturované – linearizované formě řetězců o průměrné délce 5 aminokyselin. (Přesná délka rozpoznávané oblasti peptidu je rozebrána v kapitole 1.2.2.)
Vlastní výpočet využití slovníku (*vocabulary usage*)[4] spočívá v podílu všech různých existujících slov délky *n* ku maximálnímu kombinatoricky možnému počtu slov této délky:

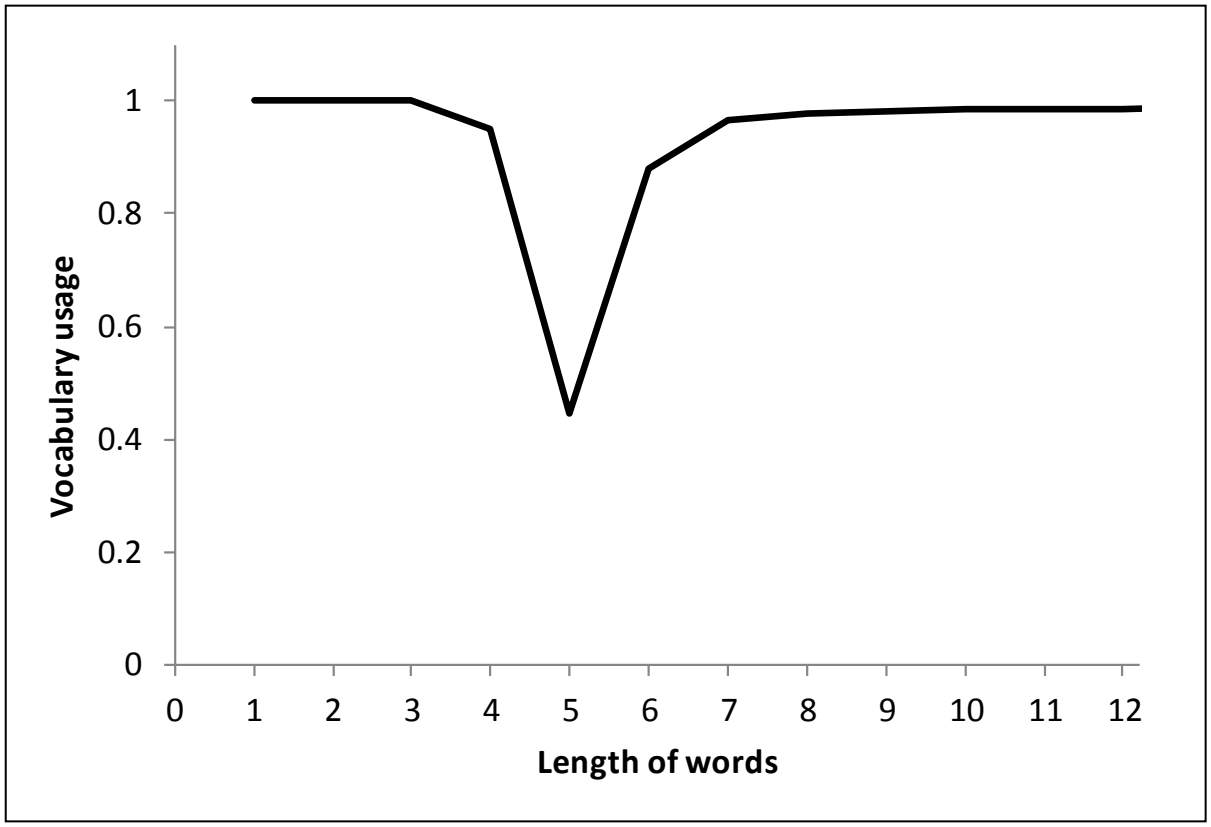
$$U_n = \frac{U_{n,a}}{U_{n,\max}}$$

Zde je třeba poznamenat, že využití slovníku má své obecné zákonitosti, které vycházejí čistě z vlastností vstupů – tedy z délky sekvence a velikosti abecedy- a nezáleží na informačním obsahu sekvence. (Platí tedy pro přirozené i náhodně generované posloupnosti.) Využití slovníku má vždy lokální minimum, jehož hodnota závisí na těchto parametrech tak, že rozsah abecedy *s* (20 aminokyselin) umocněn délkou slova *n* odpovídá řádově délce sekvence *N*. (Obr. 1) Matematicky vyjádřeno, frekvence *X* nějakého slova, které se v sekvenci vyskytuje, je náhodnou proměnnou s binomickým rozdělením o parametrech N-n+1 and $1/s^n$.

[4] Výstup našeho programu poskytuje jak tento podíl, tak celkový seznam aktuálně využitých slov a jejich frekvencí, který umožňuje následnou kvalitativní analýzu proteomu, která však již není předmětem této práce.

$$P(X=k)=\binom{N-n+1}{k}\left(\frac{1}{s^n}\right)^k\left(\frac{s^n-1}{s^n}\right)^{N-n+1-k}, \qquad k=0,1,\ldots,N-n+1.$$

Fig. 1. Využití slovníku



Podíl aktuálního a maximálně možného slovníku u náhodné sekvence o délce1 000 000 písmen při rozsahu abecedy 20 aminokyselin. Minimum je dosaženo pro délky slova 5.

Pokud je s^n vzhledem k délce celého řetězce velmi malé, je pravděpodobné, že se v řetězci vystřídají všechna, či téměř všechna, slova dané délky. Hodnota *vocabulary usage* tak bude blízká jedné. Pokud naopak počet potenciálně existujících slov bude mnohem větší, než je délka řetězce, je nepravděpodobné, že by se v něm vyskytla dvě (či více) stejná slova a řetězec tak zřejmě opět bude obsahovat maximální možný počet různých slov. Pokud je ale počet potenciálně existujících slov s^n srovnatelný s délkou řetězce, je naopak velmi pravděpodobné, že se některá slova budou opakovat a přitom nebude dost prostoru, aby se objevila všechna, jež existovat mohou. *Vocabulary usage* tak zřejmě nabude nejmenší hodnoty pro délku slova n přibližně

$$n \approx \frac{\ln N}{\ln s},$$

kde N je délka řetězce.

V případě námi studovaných sekvencí o délkách přibližně v rozmezí 1,5 – 30MB je minimum dosaženo právě pro délky slova 5. Dále je patrné, že variabilita využití slovníku je realizována pouze v rozmezí délek slov 4-6, jinak se trend blíží jedné. Jediná oblast, kde může docházet k potenciálním modifikacím, by tedy měla být v tomto rozmezí, což je přesně v souladu s naší analýzou.

2. Metodika

2.1. Sběr dat

Všechna data byla stažena z veřejně dostupných databází NCBI GenBank a Sanger Institute jako predikované proteomy ve formátu FASTA.
Výběr organismů byl omezen zejména dostupností dostatečně dlouhých proteomů v databázích. Velikost proteomu vhodného k analýze jsme zvolili empiricky na základě chování náhodně generovaných sekvencí o délce 1 000 000 aminokyselin. Vzhledem k tomu, že každý proteom ještě prošel procesem filtrace při kterém jsou eliminovány shodné nebo velmi podobné sekvence, a z tohoto, již neredundantního proteomu, byl vybrán náhodný vzorek 1 000 000 n-gramů, soustředili jsme se pouze na organismy s proteomem větším než 1,2 MB. Dále jsme se snažili o výběr takových organismů, které je možno jednoznačně zařadit jako parazitické nebo neparazitické. V případě volně žijících jednobuněčných jsme ale zařadili organismus *Tetrahymena thermophila,* přestože byl referován ojedinělý případ, že tento nálevník je schopen přejít na parasitickou životní strategii, neboť data pro volně žijící jednobuněčná eukaryota jsou vzácnější než pro parazity a bylo třeba vytvořit dostatečně velký soubor dat. Podobně v některých případech volně žijících jednobuněčných eukaryot bylo otázkou, jak klasifikovat jednobuněčnost/mnohobuněčnost v případě přechodných forem jako jsou *Dictyostelida* jež tvoří buněčné shluky (tzv. *cellular slime molds*) nebo kolonie rodu *Volvox.* Pro tyto přechodné formy jsme všechny analýzy, kde hrála roli klasifikace dle buněčnosti, udělali pro obě varianty.

Do analýzy jsme nezahrnuli prokaryota kvůli jejich zásadně odlišné evoluční historii. Rovněž jsme nezahrnuli ektoparazity a vyloučili jsme i všechny houby, neboť v případě parazitů se jedná většinou právě o ektoparazity (jako plíseň bramborová, či hmyzí patogeny z

rodu *Metarhyzium*) nebo mají příliš krátký proteom, případně vykazují příliš mnoho neobvyklostí jako „modelový organismus“ *Sacharomyces cer.*

Seznam organismů a jejich klasifikace jsou uvedeny v tabulceS2, přílohy 1.

2.2. Filtrace a standardizace dat

Pro potřeby *n*-gramové analýzy byly ze sekvencí odstraněny anotace a rovněž znaky, které nepatří do základní aminokyselinové „abecedy“ (konkrétně znaky B, J, Z, U, X) jež jsou ojediněle v sekvenci přítomny např. na místech mezer nebo neznámých aminokyselin, kde známe pouze jejich polaritu. Odstranění znaků vzhledem k málo frekventovanému výskytu nevede ke ztrátě informace, naopak jejich ponecháním by se uměle rozšířila velikost abecedy, což by zásadně měnilo výsledek lingvistické analýzy.

V proteomu se dále nachází duplicity, homology a paralogy, tedy sekvence, které se od sebe liší např. jen několika aminokyselinami. Tyto redundance ovlivňují výpočet využití slovníku. Proteomy tedy před analýzou prošly filtrací, při které zůstal vždy pouze jeden protein z rodiny podobných sekvencí. Filtrace probíhala dle následujícího algoritmu: Proteiny jsou postupně načítány a program z nich náhodně vybírá k peptidů délky n a porovnává je s již předtím načtenými proteiny. Pokud je nalezen alespoň jeden shodný peptid o daných parametrech, je vyřazen. Parametry jsou přednastaveny v programu empiricky na délku vybíraného peptidu n=16 aminokyselin a počet porovnávaných vzorků k= 5. U organismů s velkým množstvím ortologů a paralogů je potřeba nastavit parametr k na vyšší hodnoty – např. u rostlin, člověka a rodu *Trypanosoma* byl parametr nastaven na 20 až 50. Optimální nastavení je třeba kontrolovat z grafu využití slovníku, kde je zřetelně vidět, zda se pro délku vzorku 16 aminokyselin blíží hodnota využití slovníku jedničce (viz obr. 1, detailní rozebrání problematiky viz Zemková and Zahradník, 2015)
Pro všechny tyto operace a následnou *n*-gramovou analýzu jsme vyvinuli software dostupný veřejně na serveru pro sdílení vědeckých dat figshare:
http://dx.doi.org/10.6084/m9.figshare.1534499.

2.3. Analýza dat

Využití slovníku bylo spočítáno pro délky peptidů 4-12 aminokyselin z náhodného vzorku 1 000 000 *n*-gramů z každého proteomu, takže všechny porovnávané proteomy měly

stejnou délku. Horní hranice *n* byla stanovena dle konsenzu délky sestřihávaných peptidů, jež jsou vystavovány na komplexu MHC (Trombetta and Mellman).
Využití slovníku U_n je definováno dle rovnice (1) jako podíl aktuálně využitého slovníku (počet všech různých peptidů délky *n*) ku maximálnímu kombinatoricky možnému využití slovníku. Teoretický počet všech možných slov $U_{n,\max}$ byl spočítán následovně:

$$U_{n,\max} = \min\left(1\,000\,000,\, s^n\right) \tag{2}$$

Kde *n* je délka peptidu a *s* je rozsah abecedy, tedy 20 aminokyselin.
Veškeré výpočty jsme dělali nejprve pro redundantní data bez filtrace a poté pro filtrovaná data. (V případě nefiltrovaných dat se statistické rozdíly mezi skupinami organismů neprojevily.) Konkrétní hodnoty využití slovníku pro délky peptidů 4-12 aminokyselin z filtrovaných dat jsou uvedeny v tabulce S3 přílohy 1.

Abychom zredukovali počet devíti závislých proměnných (U_4-U_{12}) a zjistili, zda v datovém souboru jsou skryté trendy vyjádřitelné pomocí nezávislých proměnných, aplikovali jsme analýzu hlavních komponent (*principle component analysis*, dále PCA), pomocí které je ve vícerozměrném prostoru převedena původní hodnota U_n na tzv. Komponentní skore (*component score* – CS). Takto jsme zjistili, že datový soubor lze charakterizovat pomocí 4 nezávislých proměnných – tzv. hlavních komponent (PC) z nichž každá vysvětlovala více než 1% variability. Dohromady tyto čtyři komponenty vysvětlují 99,9% variability. Pro výpočet jsme použili standardní balíček softwaru R – PCA z kovarianční matice, bez rotace) (*R Core Team, 2014).*

Pomocí analýzy kovariance jsme zjistili, jakou měrou ovlivňují jednotlivé faktory konkrétní hlavní komponenty (ANCOVA, suma čtverců III. typu) Jako kovariáty byly vzaty 4 charakteristiky proteomu jednotlivých organismů: délka proteomu před filtrací, délka po filtraci, počet proteinů před filtrací a po filtraci. Vzhledem k hierarchické struktuře dat (tedy např. uvnitř datasetu neexistuje autotrofní organismus, který by byl zároveň parazit) jsme nemohli všechny organismy zahrnout do jednoho komplexního modelu, namísto toho jsme udělali pro každou hlavní proměnnou *Z* (parazitismus, jednobuněčný parazitismus, mnohobuněčný parazitismus, mnohobuněčnost, heterotrofie a extracelulární parazitismus) oddělený model obsahující vedlejší proměnné $X_1 - X_4$:

$$CS_i = \beta_0 Z_i + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \varepsilon_i, \tag{2}$$

Kde CS_i je konkrétní komponentní skore *i*-tého organismu, $X_{1,i}$ je délka nefiltrovaného (redundantního) proteomu, $X_{2,i}$ je délka filtrovaného proteomu, $X_{3,i}$ je počet proteinů nefiltrovaného proteomu, $X_{4,i}$ je počet proteinů filtrovaného proteomu, b_0 – b_4 jsou regresní koeficienty a ε_i je chybový člen. *Z* je hlavní proměnná – identifikátor binárního charakteru, který je stanoven podle toho, zda konkrétní organismus je:

-Parazit/volně žijící

-Jednobuněčný parazit/ volně žijící jednobuněčný

-Jednobuněčný/mnohobuněčný parazit

-Extracelulární/intracelulární parazit

-Jednobuněčný/mnohobuněčný

-Autotrof/heterotrof.

Do původní analýzy byly zahrnuty i proměnné reflektující příslušnost k obratlovcům a k říši Excavata. V obou případech nebyly žádné významné rozdíly nalezeny a do výsledné analýzy nebyly zahrnuty zejména kvůli malému počtu dat.

3. Výsledky a diskuse

Pro 71 organismů byly spočteny hodnoty využití slovníku pro délky peptidů od 4 do 12 aminokyselin. Hodnoty pro filtrovaná data jsou uvedeny v tabulce S3 přílohy 1Již z těchto dat je patrné, že největší rozdíly v individuálních hodnotách jsou na úrovni délek peptidů 4-6 aminokyselin a dále trend směřuje v podstatě uniformně k jedničce[5]. Dle analýzy hlavních komponent jsme zjistili, že 4 hlavní komponenty vysvětlují 99,9% variability využití slovníku. Faktory ovlivňující jednotlivé komponenty jsou jednotlivé délky peptidů. Sycení jednotlivých komponent konkrétními faktory (*factor loading*) poukazuje na základní trendy ve způsobu využití slovníků (Obr. 1 přílohy 1), jež lze shrnout následovně:

[5] Pokud naopak pozorujeme u delších peptidů jednotlivých organismů odchylky a u některého organismu jsou hodnoty pro U_7-U_{12} nižší, je to znamení, že proteom nebyl dostatečně vyfiltrován, nebo se jedná o špatně anotovanou sekvenci (diverzita slovníku pro tyto délky je nižší, tedy pravděpodobně snížená existencí repetic). V takovém případě je třeba provést filtraci znovu případně zkontrolovat celý proteom a jeho webový zdroj.

Hlavní komponenta	% vysvětlené variability	Převažující trend U_i
PC1	**59,9**	Výrazně snížená diverzita pentapeptidů Výrazně zvýšená diverzita hexapeptidů
PC2	**34,2**	Výrazně snížená diverzita pentapeptidů a hexapeptidů Obecná chudost celého slovníku z výjimkou slabě nabohacených tetrapeptidů
PC3	3,9	Obecná chudost slovníku s výjimkou slabě nabohacených pentapeptidů a hexapeptidů
PC4	2,9	Výrazně zvýšená diverzita tetrapeptidů, slabě zvýšená diverzita hexapeptidů Obecná chudost slovníku

Na využití slovníku mohou mít vliv kromě ekologických faktorů jako parazitismus a autotrofie také fylogenetické faktory a faktory charakterizující proteom (délka, redundance, počet proteinů atd.)

Pomocí analýzy kovariance jsme zjistili, že největší příspěvek k vysokým hodnotám PC1 mají parazité a to jak jednobuněční, tak mnohobuněční. K PC2 obecně přispívají jednobuněčné organismy. Přestože třetí a čtvrtá komponenta vysvětlují mnohem méně variability než první dvě komponenty, zabýváme se s nimi ve výsledcích proto, že pravděpodobně reflektují rozdíl mezi jednobuněčným a mnohobuněčným parazitismem (jednobuněční parazité přispívají kladně k PC3) a rozdíl mezi intracelulárním a extracelulárním parazitismem (extracelulární parazité přispívají kladně k PC4). Efekt parazitismu převažuje i nad efektem délky filtrovaného proteomu, který je druhým nejsilnějším ovlivňující faktorem PC1 (viz tabulka S4 přílohy 1- kompletní přehled analýzy kovariance). Ostatní faktory související s kvalitou proteomu neměly na hodnoty hlavních komponent vliv[6]. Z uvedeného vyplývá, že parazité mají statisticky významně sníženou diverzitu pentapeptidů a naopak zvýšenou diverzitu hexapeptidů. Jednobuněční (parazité i neparazité) mají obecně chudší slovník**.** Extracelulární parazité mají výrazně zvýšenou diverzitu tetrapeptidů. Vlastní postavení organismů lze nejlépe pozorovat v dvourozměrném prostoru definovaném jednotlivými komponentami. Pro hypotézu o vlivu parazitismu na využití slovníku je nejdůležitější komponentní skóre PC1 a PC3 (obr. 2 přílohy 1) – tedy pozice v dvourozměrném prostoru definovaném komponentami, na něž má parazitismus signifikantní vliv. Parazité zde tvoří klastr v oblasti kladných hodnot PC1 (pravá strana grafu), volně žijící organismy se nacházejí převážně v oblasti záporných hodnot PC1 (levá strana grafu). Rozdíl mezi jednobuněčnými a mnohobuněčnými je nejlépe patrný z komponentního skóre PC1 a PC2(obr. S1 přílohy 1) Mnohobuněční jasně okupují oblast záporných hodnot

[6] (S výjimkou počtu proteinů nefiltrovaného proteomu v případě příspěvku jednobuněčných parazitů k PC3, viz tabulka S4 přílohy 1, kde je tato anomálie způsobena početnou skupinou *Leishmanií*, které mají relativně vysoké množství homologů a paralogů. Je otázkou, zda je tato vlastnost přirozená, nebo uměle způsobená např. specifiky technologií, kterými jsou sekvence zpracovány.)

PC2 (dolní část grafu)[7]. Z komponentního skóre PC1 a PC4 (Obr S2 přílohy 1) lze vyčíst rozdíly mezi extracelulárními a intracelulárními parazity. (Extracelulární parazité okupují převážně I. Kvadrant – tedy oblast kladných hodnot PC1 i PC4.)
Abychom vyloučili fylogenetické ovlivnění výsledku, testovali jsme, jak se liší slovníky parazitů a neparazitů ze stejné fylogenetické skupiny. Takových skupin jsme měli pouze 5 (Kinetoplastida, Ciliates, Nematoda, Opisthokonta a Sar), ale ve všech skupinách byla průměrná hodnota PC1 parazitů vždy vyšší než u volně žijících. Stejně tak v případě PC3 měli parazité vždy nižší hodnoty PC3 než volně žijící – tedy v souladu s hypotézou. (Statistická významnost rozdílů byla ověřena pomocí exaktní varianty binomického testu.)

3. 1. MHC-hypotéza

Výsledky ukazují, že k modifikaci peptidického slovníku dochází v rozmezí délek peptidů 4-6 aminokyselin. To je v souladu s přirozenými kombinatorickými vlastnostmi sekvencí o rozsahu abecedy 20 písmen a délkách sekvencí přibližně od 1,5 – 30MB (viz kapitola 1.3.2.)

Princip imunitního rozpoznávání antigenu pomocí T-receptoru je podrobně popsán v úvodu v kapitole 1.2. Ukazuje se, že přestože délka sestřihávaných peptidů uvnitř buňky je v rozmezí zhruba 8-13 aminokyselin, na vlastním imunitním rozpoznání se podílí v případě MHC I řetězec o délce 4-5 aminokyselin a ostatní slouží k ukotvení uvnitř molekuly MHC a zprostředkování vazby na T-receptor. Snížení diverzity pentapeptidů (a v nižší míře i tetrapeptidů) u parazitů by se tedy jevilo jako logické. Inverzní zvýšení diverzity hexapeptidů by bylo možné brát jako prostředek k zachování jakési „minimální lingvistické komplexity“ potřebné k uchování všech důležitých funkcí organismu.

V případě komplexu MHC II, kde jsou rozpoznávány primárně extracelulární agens je délka rozpoznávaného peptidu delší, protože vazebné místo je, na rozdíl od MHC I, otevřené na obou koncích. (Délku lze odhadnout z velikosti efektorové oblasti tzv. CLIP fragmentu – v případě lidských buněk dlouhého 9 aminokyselin.) Vzhledem k tomu, že na MHC I jsou presentovány peptidy převážně z vnitřního prostředí buňky, zatímco na MHC II spíše

[7] Zde je patrná skupina organismů s výrazně nízkými hodnotami PC2 (tedy nápadně vysokou diverzitou slovníku) vyznačená v obr. elipsou. Jedná se ve všech případech o organismy s poměrně krátkým proteomem o délkách 1,2MB -1,7MB. Tyto organismy vytvořily oddělený klastr, protože jejich využití slovníku je relativně bohatší v porovnání s organismy s delším proteomem. Přestože jejich aktuálně využitý slovník je chudší, zabírá větší část kombinatoricky možného slovníku, takže jejich využití slovníku (jak je definováno v rovnici 1) je vyšší. Přestože náš postup zde narazil na délkové limity ovlivňující výsledek využití slovníku, ponechali jsme tyto organismy v našem souboru, neboť jejich odstraněním bychom přišli o významná data a jejich přítomnost nemá vliv na hlavní výsledek platný pro první hlavní komponentu spojený s parazitismem. V případě opakování této nebo podobné analýzy v budoucnu však je třeba tuto limitaci vzít v potaz.

z vnějšího prostředí, dalo by se očekávat, že se budou lišit ve využití slovníku extracelulární a intracelulární parazité. Rozdíl byl skutečně pozorován, ale pouze v případě PC4 (vysvětlující pouze 2,9% variability). Extracelulární parazité mají, dle našeho výsledku výrazně zvýšenou diverzitu tetrapeptidů a o něco méně zvýšenou diverzitu hexapeptidů, zatímco zbytek slovníku je celkově ochuzený. Je však třeba si uvědomit, že samotné dělení na intracelulární a extracelulární parazity je problematické, protože někteří parazité (např. *Trypanosoma cruzi*) během života mohou vystřídat obě prostředí. Někteří paraziti (jako četní příslušníci kmene Apikomplexa) si budují uvnitř hostitelské buňky tzv. parazitoforní vakuolu[8].

Podobně neostrá je hranice mezi prezentací čistě extracelulárních a intracelulárních peptidů, protože je znám i proces tzv. cross-prezentace, kdy je dendritickými buňkami degradováno extracelulární agens a poté přeneseno dovnitř buňky, kde je zpracováno již standardní cestou.

Problémem MHC –hypotézy je její obtížná testovatelnost na protipříkladech parazitických organismů, které parazitují na hostitelích, kteří nemají imunitní systém založený na MHC. V našem souboru figuruje z jednobuněčných pouze parazit ústřice – *Perkinsus marinus.* Ten skutečně má, na rozdíl od parazitů, výrazně zvýšenou diverzitu pentapeptidů (a zachované – pro parazity typické zvýšení hexapeptidů). Dalším zástupcem v našem souboru je pouze mnohobuněčný hmyzí patogen *Heterohabditis bacteriophora,* který se však chová jako typický mnohobuněčný parazit (nachází se ve IV. Kvadrantu komponentního skóre PC1 a PC3 – tedy má kladné hodnoty PC1 a záporné hodnoty PC3 v souladu s většinou mnohobuněčných parazitů. Další organismy odpovídající podmínkám stanovených pro náš pokus se nám bohužel nepodařilo nalézt. Navíc absence MHC –imunitního rozpoznávání neznamená absenci imunity obecně a ukazuje se, že rozpoznávací mechanismy jsou založeny na obecných principech sdílených napříč říšemi, jejichž součástí můžou být i preferované délky peptidů vycházející ze stejných kombinatorických důvodů jako v případě MHC rozpoznávacího systému.

3. 2. Výjimky z trendu

Kromě již zmíněného *P. marina* je vidět v komponentním skóre PC1 a PC3 následující výrazné výjimky: *Ostreococcus tauri,* drobná zelená řasa známá svou celkovou funkční redukovaností, který vykazuje trend slovníku odpovídající parazitickému chování, pro které

[8] Není bez zajímavosti, že právě určitá skupina apikomplex tvořící specifický typ parazitoforní vakuoly sdílí typickou modifikaci peptidového slovníku ochuzení jak pentapeptidů, tak hexapeptidů.

nemáme žádné uspokojivé vysvětlení. Opačný trend naopak pozorujeme u parazita *Clonorchis sinensis,* který je sice v souladu s mnohobuněčnými organismy, ale má záporné hodnoty PC1. V případě tohoto organismu můžeme hledat vysvětlení např. v možné kontaminaci nebo v chybách vzniklých při predikci proteomu. Pro srovnání jsme použili i starší verzi proteomu založenou na predikci z transkriptomu a pozice *C. sinensis* se posunula výrazně blíže ke klastru parazitů. Přestože kladných hodnot PC1 nebylo dosaženo, odchylka z trendu by již nebyla zdaleka tak výrazná (Viz obr 1 přílohy 1).

V případě volně žijících organismů *N. gaditana, C. elegans, Ph.tricornutum* a *O. sativa* je posun směrem ke kladným hodnotám PC1 způsoben délkou filtrovaného proteomu na dolní hranici limitu.

Nepatogenní *Entamoeba dispar* žijící jako komenzál v trávicím traktu užívá z hlediska strategie stejné krycí mechanismy jako *E. Histolytica,* není tedy překvapením, že se její využití slovníku od její parazitické příbuzné neliší.

4. Závěr

V naší studii byla testována apriorní hypotéza o redukci peptidického slovníku parazitických organismů ve srovnání s organismy volně žijícími. Tuto hypotézu potvrdily čtyři nezávislé důkazy. Konkrétně ochuzený slovník jednobuněčných parazitů, mnohobuněčných parazitů, výsledek kontrastního testu u pěti párů sesterských taxonů obsahujících zároveň parazitické i neparazitické druhy a fakt, že *Perkinsus m.* (jeden ze dvou parazitů na hostitelích postrádající imunitní systém založený na MHC) má neredukovaný slovník srovnatelný bohatostí se slovníky volně žijících jednobuněčných. Problémem studie zůstává nedostatek protipříkladů – tedy adekvátních dvojic parazitů a hostitelů postrádajících imunitní systém založený na MHC. (Dostupní četní parazité rostlin bohužel nesplňují kritéria délky proteomu, nebo jsou ektoparazité.)

Přestože to nebylo cílem studie, výsledky ukázaly, že mnohobuněčné organismy mají obecně bohatší pentapeptidové i hexapeptidové slovníky (přispívají záporně k PC2), což může být způsobeno celkově jejich vyšší komplexitou ve srovnání s jednobuněčnými. Nicméně je třeba poukázat na to, že do studie byly zahrnuty pouze reprezentanti 3 větví mnohobuněčných (Metazoa, Metaphyta a Charophyta).

Z výsledků vyplývá, že T-buňky imunitního systému rozpoznávají peptidy o délce 4-5 aminokyselin, nikoliv celý peptid vázaný na MHC I, což může být důležité pro design vakcín.

Analýza byla provedena na datech dostupných v květnu 2015. V budoucnu bude jistě možné tento soubor rozšířit a revidovat dle nově dostupných anotací a provést tak nezávislou kontrolu studie. Pomocí softwaru, který byl vytvořen pro potřeby této analýzy, by bylo rovněž možné testovat hypotézy o podobnosti slovníků mezi parazitem a jeho specifickým hostitelem.

5. Literatura

Austyn, J. M., and K. J. Wood, 1994, Principles of Cellular and Molecular Immunology: New York, Oxford University Press.

Beck, S., and J. Trowsdale, 2000, The human major histocompatability complex: lessons from the DNA sequence: Annu Rev Genomics Hum Genet, v. 1, p. 117-37.

Berzofsky, J., 1989, Structural features of T-cell recognition: applications to vaccine design: Philosophical Transactions of the Royal Society of London. B, Biological Sciences, v. 323, p. 535-544.

Boehm, T., and J. B. Swann, 2014, Origin and evolution of adaptive immunity: Annu. Rev. Anim. Biosci., v. 2, p. 259-283.

Bolshoy, A., 2003, DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity: Appl Bioinformatics, v. 2, p. 103-12.

Brendel, V., J. S. Beckmann, and E. N. Trifonov, 1986, Linguistics of nucleotide sequences: morphology and comparison of vocabularies: Journal of Biomolecular Structure and Dynamics, v. 4, p. 11-21.

Cvrčková, F., 2006, Úvod do praktické bioinformatiky, Academia.

Chang, S. C., F. Momburg, N. Bhutani, and A. L. Goldberg, 2005, The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a "molecular ruler" mechanism: Proc Natl Acad Sci U S A, v. 102, p. 17107-12.

Hořejší, V., J. Bartůňková, T. Brdička, and R. Špíšek, 2013, Základy imunologie, Triton.

Karlin, S., I. Ladunga, and B. E. Blaisdell, 1994, Heterogeneity of genomes: measures and values: Proceedings of the National Academy of Sciences, v. 91, p. 12837-12841.

Kasahara, M., J. Nakaya, Y. Satta, and N. Takahata, 1997, Chromosomal duplication and the emergence of the adaptive immune system: Trends Genet, v. 13, p. 90-2.

Kasahara, M., T. Suzuki, and L. D. Pasquier, 2004, On the origins of the adaptive immune system: novel insights from invertebrates and cold-blooded vertebrates: Trends Immunol, v. 25, p. 105-11.

Kelley, J., L. Walter, and J. Trowsdale, 2005, Comparative genomics of major histocompatibility complexes: Immunogenetics, v. 56, p. 683-95.

Kopecký, J., and O. Kopecký, 2010, NK buňky, chemokiny a chemokinové receptory: Klinická onkologie, v. 1, p. 5-7.

Kürn, U., F. Sommer, G. Hemmrich, T. C. Bosch, and K. Khalturin, 2007, Allorecognition in urochordates: identification of a highly variable complement receptor-like protein expressed in follicle cells of Ciona: Dev Comp Immunol, v. 31, p. 360-71.

Madden, D. R., 1995, The three-dimensional structure of peptide-MHC complexes: Annual review of immunology, v. 13, p. 587-622.

Neefjes, J., and H. Ovaa, 2013, A peptide's perspective on antigen presentation to the immune system: Nat Chem Biol, v. 9, p. 769-75.

Orlov, Y. L., and V. N. Potapov, 2004, Complexity: an internet resource for analysis of DNA sequence complexity: Nucleic acids research, v. 32, p. W628-W633.

Paz, P., N. Brouwenstijn, R. Perry, and N. Shastri, 1999, Discrete Proteolytic Intermediates in the MHC Class I Antigen Processing Pathway and MHC I–Dependent Peptide Trimming in the ER: Immunity, v. 11, p. 241-251.

Pietrokovski, S., J. Hirshon, and E. Trifonov, 1990, Linguistic measure of taxonomic and functional relatedness of nucleotide sequences: Journal of Biomolecular Structure and Dynamics, v. 7, p. 1251-1268.

Popov, O., D. Segal, and E. N. Trifonov, 1996, Linguistic complexity of protein sequences as compared to texts of human languages: Biosystems, v. 38, p. 65-74.

Rammensee, H. G., T. Friede, and S. Stevanoviíc, 1995, MHC ligands and peptide motifs: first listing: Immunogenetics, v. 41, p. 178-228.

Ronald, P. C., and B. Beutler, 2010, Plant and animal sensors of conserved microbial signatures: Science, v. 330, p. 1061-4.

Shannon, C. E., 1948, A mathematical theory of communication: The Bell System Technical Journal, v. 27, p. 379–423, 623–656.

Sharon, I., A. Birkland, K. Chang, R. El-Yaniv, and G. Yona, 2005, Correcting BLAST e-values for low-complexity segments: Journal of computational biology, v. 12, p. 980-1003.

Trombetta, E. S., and I. Mellman, 2005, Cell biology of antigen processing in vitro and in vivo: Annu. Rev. Immunol., v. 23, p. 975-1028.

Vogt, A., and H. Kropshofer, 2011, CLIP - a multifunctional MHC class II associated self - peptide, *in* A. Kastin, and A. J. Kastin, eds., Handbook of Biologically Active Peptides, Academic Press, p. 611-621.

Vyas, J. M., A. G. Van der Veen, and H. L. Ploegh, 2008, The known unknowns of antigen processing and presentation: Nature Reviews Immunology, v. 8, p. 607-618.

Wolfe, K. H., 2001, Yesterday's polyploids and the mystery of diploidization: Nature Reviews Genetics, v. 2, p. 333-341.

Zemková, M., Zahradník, D., 2015 Complexity- Software tools for linguistic based analysis of genetic sequences.Figshare. http://dx.doi.org/10.6084/m9.figshare.1534499.

One common structural feature of “words” in protein sequences and human texts

1. Úvod

V článku je ilustrována strukturní analogie mezi sekvencemi aminokyselin tvořící amfipatický alfa helix a lidskými slovy. V případě struktury amfipatického helixu se střídají ve specifické periodě polární (P) a nepolární (N) aminokyseliny. V lidských promluvách pak alternují podobně souhlásky a samohlásky. Tato studie nepředpokládá jakoukoliv spojitost mezi lidskými a genetickými texty. Na příkladu 32 proteomů různých organismů a dvou typů lidských textů (poetického a technického) ve třech indoevropských jazycích jsou demonstrovány alternace polárních a nepolárních aminokyselin a analogicky střídání souhlásek a samohlásek. Nejde zde o porovnávání lidských a genetických textů co do jejich matematických charakteristik. Takový postup by neměl oprávnění už jen z toho důvodu, že jak sekvence proteinu, tak psaná podoba lidské promluvy jsou pouze modely fyzického světa převedené na lineární posloupnosti prvků abecedy. Přestože takové pokusy tu již byly (Popov et al. 1996), je diskutabilní, co je vlastně v obou sekvencích „znakem“ ve smyslu nesení významu. V případě lidských jazyků je to patrně „něco mezi slovem a slabikou“, nejspíše část slova označovaná jako morfém (nejmenší jednotka jazyka nesoucí význam). Tu ovšem lze jen velmi těžko algoritmicky extrahovat. I kdyby se to podařilo, nemělo by smysl porovnávat kvantitativně morfémy a například oligopeptidy, protože není žádný důvod se domnívat, že mezi sekvencemi lidských a genetických textů existuje nějaká spojitost. Pozorujeme-li strukturně nějakou „podobnost“, jedná se o analogii vycházející jednak ze skutečnosti, že 1) v případě obou typů sekvencí jde o lidský konstrukt fyzického světa uvedený do linearizované a digitalizované podoby. 2) V obou případech se informační struktura schopná nést význam, vytvořila díky existenci několika málo vnějších okrajových podmínek, které navíc lze diskrétně převést na stav 0/1. Dodejme ovšem, že jak v případě polarity aminokyselin, tak v případě rozčlenění hlásek je digitalizace výsledkem konsenzu, nikoliv úplně jejich přirozené povahy. (Některé aminokyseliny se mohou chovat v prostředí okolních aminokyselin jinak co do polarity, podobně některé souhlásky jako např. slabikotvorné L a R ve slovanských jazycích mohou fungovat jako samohlásky.)

1.2. Rozklad textu na shannonovské n-gramy

Jak již bylo podrobněji rozebráno v kapitole 1.3.2. předchozího článku, *n*-gramy jsou, v případě rozkladu sekvence proteinu, posloupnosti *n* aminokyselin. Aminokyseliny rozdělujeme na *polární* (P) a *nepolární* (N). Každou možnou posloupnost polárních a nepolárních aminokyselin (např. PNNN, PNNP) budeme nazývat *typ n-gramu*. Některé z těchto typů jsou *amfipatické*, ostatní jsou *neamfipatické*. *Unikátní n-gramy* budeme nazývat soubor navzájem odlišných n-gramů. Úplný seznam všech amfipatických a neamfipatických 12-gramů čítá při dvou písmenné abecedě (P,N) vždy 2^{12} (4096) – z toho unikátních amfipatických 12-gramů 896 a unikátních neamfipatických 12-gramů 3200.

1. 3. Struktura amfipatického helixu

Amfipatický alfa helix je spolu se strukturou beta-sheet nejčastější sekundární proteinovou strukturou. V proteomu jsou kromě hlavních sekundárních struktur přítomny i peptidy, jež nejsou strukturované vůbec, anebo jen částečně, případně nabydou struktury teprve spojením s jinými proteiny.(Dunker et al. 2000; Dunker et al. 2008) Peptidy, které by měly potenciálně tvořit amfipatický alfa helix je možné identifikovat v sekvenci na základě typické periody 3,5 [9]mezi polárními (hydrofilními) a nepolárními (hydrofobními) aminokyselinami (tabulka 1 přílohy 2) (Mant et al. 1993). To znamená, že vzdálenost mezi aminokyselinami stejné povahy je 3 nebo 4. V takovém případě je jedna strana helixu polární, zatímco opačná nepolární. Typická délka helixu je 10- 15 aminokyselin (Berezovsky et al. 2000).

1. 4. Souhlásky a samohlásky v lidských jazycích

Množství souhlásek a samohlásek (*consonant/vowel inventory)* v různých jazycích se podstatně mění, stejně tak jejich vzájemný poměr. Extrémními případy konsonantálních jazyků jsou berberské jazyky, které užívají až kolem stovky různých konsonant, příkladem jazyků, kde jsou využívány vokály je skupina Havajských jazyků. Existuje celá teorie a členění jazyků dle poměru souhlásek a samohlásek včetně spekulací, zda souvisejí nějak s komplexitou jazyka (Lehmann 1992) (Maddieson, 2013). Ukazuje se ale, že všechny jazyky dosahují v podstatě stejné komplexity, jen různými prostředky a kvantitativní analýzy poměrů a počtů hlásek nám ve skutečnosti o daném jazyku mnoho neřeknou. Díky tomu bereme celý

[9] Perioda se pohybuje v intervalu (3,1; 3,7) –liší se u různých organismů – především mezi prokaryoty a eukaryoty.

model jako čistě ilustrativní a pro tyto potřeby zjednodušený. Hlásky jsou zde z praktických důvodů ztotožněny s písmeny. Celá problematika zvukové rozmanitosti v různých jazycích je pominuta. Stejně tak skutečnost, že dokonalý model, kde je zastoupeno více různých jazyků, by vyžadoval jednotnou fonetickou transkripci, což je v podstatě nesplnitelný požadavek.

2. Metodika

2. 1. Analýza proteinů

2. 1. 1. Zdrojová data

Celkově bylo analyzováno 32 proteomů, jež byly staženy ve formátu FASTA z veřejné databáze NCBI. Detailně jsme pak věnovali pozornost deseti organismům, které byly vybrány tak, aby pokryli co nejširší spektrum variability (Tabulka 2 přílohy 2).
Z proteomů byly odstraněny homology a paralogy (proteiny lišící se jen v několika málo aminokyselinách) Použitý algoritmus je již detailně popsán v článku Reduction of peptide vocabulary in parasites – an evolutionary immune evasion strategy? (Zemková et al., 2016)
Pro všechny operace s *n*-gramy včetně statistického testu byl vyvinut software „Complexity_G“, dostupný nyní z Figshare.

2. 1. 2. *N*-gramová analýza

Z každého proteomu byl vybrán náhodný vzorek 1 000 000 12-gramů, tímto jsme vytvořili slovník peptidů o délce 12 aminokyselin. Délka *n*-gramu byla stanovena na základě konsenzu pro délku amfipatického alfa helixu (Berezovsky et al. 2000). Seznam 12-gramů byl seřazen sestupně dle frekvence. (Frekvence jednotlivých n-gramů je automatickým výstupem programu.) Sekvence aminokyselin byly konvertovány na posloupnosti P/N dle převodní tabulky 1 přílohy 2.

Jednotlivé 12-gramy v seznamu byly označeny jako amfipatické nebo neamfipatické dle následujícího algoritmu:
Sekvence (P,N) tvoří amfipatický alfa - helix jestliže P, PP, nebo PPP alternuje s N, NN, nebo NNN . Tedy že existuje P (tzv. koordináta 0), a na 3. nebo 4. pozici (případně na obou)následuje další P a stejně tak na pozici 7., 10. nebo 11, případně obou. To samé platí i pro N. (Z tohoto vyplývá, že nejmenší možná délka pro detekci potenciálně amfipatické sekvence je 4, protože ta potenciálně obsahuje alespoň jednu otočku amfipatického alfa-helixu a může být součástí delšího helixu o několika periodách.)

Kompletní seznam všech 4096 12-gramů byl pracovně rozdělen na 10 částí po 400 12-gramech (zbylých 96 12-gramů nebylo bráno v potaz, neboť v koncových částech seznamu se

amfipatické peptidy stejně již téměř nevyskytují) V každé části byl spočítán počet unikátních amfipatických 12-gramů a zaznamenán do grafu. Takto jsme získali distribuční křivku amfipatických 12-gramů v proteomu (Obr 1 přílohy 2). Je-li trend distribuční křivky sestupný (tedy více amfipatických n-gramů se nachází ve vrchní části seznamu), pak jsou amfipatické peptidy preferovány.

2. 1. 3. Statistické vyhodnocení

Mějme text rozdělený do posloupnosti *n*-gramů. Jednotlivé n-gramy jsou v textu zpravidla zastoupeny s rozdílnou četností a v posloupnosti je proto uspořádáme podle četnosti v sestupném pořadí. Podle vnitřního uspořádání rozlišujeme 2 typy *n*-gramů: amfipatické a neamfipatické. Předpokládá se, že amfipatické n-gramy jsou preferovány, tj. v uspořádané posloupnosti n-gramů se vyskytují s větší pravděpodobností v její horní části. K ověření této domněnky jsme použili Jednovýběrový Kolmogorovův – Smirnovův test, pomocí kterého jsme ověřili, že *n*-gramy nejsou rozloženy rovnoměrně. Z grafu empirické distribuční funkce lze následně vyčíst, zda je nerovnoměrnost způsobena větší pravděpodobností výskytu amfipatických *n*-gramů na začátku posloupnosti nebo v jiných částech seznamu. (Fig. 1).

Test spočívá ve vyhodnocení rozmístění amfipatických *n*-gramů mezi všemi n-gramy. Označme X pořadové číslo náhodně vybraného amfipatického *n*-gramu v uspořádané posloupnosti všech *n*-gramů. Pokud by amfipatické n-gramy nebyly preferovány, byla by stejná pravděpodobnost jejich výskytu v libovolné části posloupnosti, tj. veličina X by měla rovnoměrné rozdělení pravděpodobnosti. Označme F empirickou distribuční funkci náhodné veličiny X a G distribuční funkci rovnoměrného rozdělení s odpovídajícími parametry. Testovací statistika pro Kolmogorovův – Smirnovův test je

$$D = \sup_x |F(x) - G(x)|.$$

Nulovou hypotézu (o rovnoměrnosti rozmístění amfipatických n-gramů) zamítneme, pokud D překročí kritickou hodnotu $D_n(\alpha)$ (Anděl, 1998; (Sheskin et al. 2011). Přesné kritické hodnoty lze pro velké n aproximovat pomocí vztahu

$$D_n(\alpha) \doteq \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$$

(Likeš and Laga,1978).

2. 2. Analýza lidských textů

Pro ilustrační potřeby jsme užili dva texty odlišných žánrů ve třech indoevropských jazycích reprezentujících větve románské, anglosaské a slovanské: 1) Román Muž, který sázel stromy, jehož autorem je Jean Giono (francouzský originál v překladu do češtiny a angličtiny) 2) Text deklarace lidských práv v těchto jazycích. Pro potřeby analýzy byly ignorovány mezery mezi písmeny. Jednotlivé hlásky byly konvertovány na posloupnosti souhlásek a samohlásek (V,C) Byly vytvořeny seznamy slov o délkách 4-6 písmen v jednotlivých jazycích s příslušnými frekvencemi. Jako alternující slova (obdobná amfipatickým sekvencím) jsou zde brána taková, kde se přímo střídají C a V, případně CC a VV – tedy jsou vyloučeny dlouhé sledy samohlásek nebo souhlásek. Je třeba poznamenat, že odstraněním mezer se do seznamu slov dostanou i slova neexistující, ve skutečnosti vzniklá uměle spojením dvou sousedních slov (sekvence C,V uvedené v seznamech nemusí vždy korespondovat s existujícími slovy). Na rozložení hlásek a na faktu, že nelze vyslovit dlouhý sled samohlásek nebo naopak souhlásek, to ale nic nemění, tudíž můžeme tuto skutečnost zanedbat. Pro všechny operace s *n*-gramy byl vyvinut software Complexity_H, který je veřejně dostupný ke stažení.

3. Výsledky a diskuse

Dle výsledků tvoří amfipatické peptidy 20-24% všech 12-gramů v proteomech námi studovaných organismů. U lidských jazyků je podíl alternujících slov vyšší (57-89%). V horní polovině seznamu ale v obou případech alternující struktury dominují (58 - 74% v případě amfipatických struktur, 90-100% v případě alternujících lidských slov).

3. 1. Obecné vlastnosti distribučních křivek amfipatických peptidů

Po analýze zatím studovaných proteomů se ukazuje, že trend distribuční křivky amfipatických peptidů je vždy sestupný – tedy více amfipatických peptidů se v sestupně seřazené posloupnosti 12-gramů vyskytuje na začátku než na konci – jinými slovy, amfipatické peptidy jsou preferovány a to i přesto, že jejich absolutní počet v rámci všech 12-gramů dané posloupnosti je menší.

Dalším typickým jevem všech distribučních křivek je fluktuace počátku křivky. Jak již bylo zmíněno výše, s největší četností se vyskytují v proteomu řetězce čistě polárních nebo

nepolárních aminokyselin (tedy neamfipatické řetězce), které pravděpodobně svědčí o vysokém podílu přirozeně nestrukturovaných (*natively unfolded* nebo též *intrinsicaly disordered proteins*) – Viz následující příklad:

Př.: Prvních sedm 12-gramů v posloupnosti proteomu organismu *Monosiga brevicolis*

NNNNNNNNNNN	3318
PPPPPPPPPPP	2996
PNPNPNPNPNPN	1587
NPNPNPNPNPNP	1555
NNNNNNNNNNNP	888
PNNNNNNNNNNN	807
NPNNNNNNNNNN	714

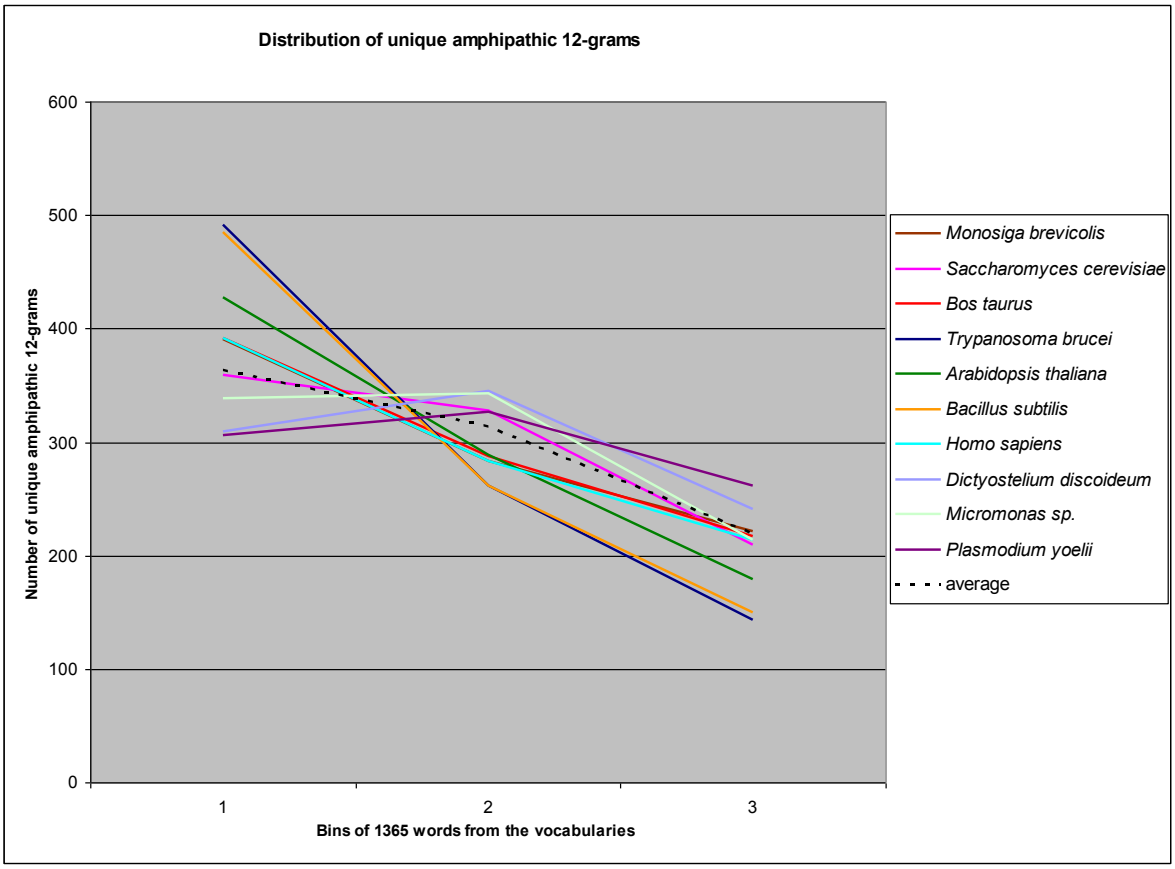
.
.
.

Počátek křivky je tedy také určitou individuální charakteristikou jednotlivých zkoumaných druhů. Zajímavým a očekávaným výsledkem je zde pozorování, že prokaryontní organismy, které mají dle současných poznatků až o třetinu menší podíl přirozeně nestrukturovaných proteinů a též zkoumané proteomy virů se vyznačují větším množstvím amfipatických peptidů na samém počátku křivky než většina ostatních organismů (tedy mají méně neamfipatických na počátku posloupnosti) (Obr. 1 přílohy 1).

Třetí vlastnost bychom mohli označit jako plynulost nebo kolísavost křivky. U studovaných organismů lze vylišit proteomy s víceméně plynulým sestupným trendem. Záleží na tom, jak jemné měřítko k pozorování si zvolíme. Máme- li proteom rozdělen pouze na 3 intervaly, pak se zde projevují dva trendy: a) Křivka je **konkávní** (Maximum unikátních amfipatických 12-gramů neleží na počátku posloupnosti, ale v její druhé třetině, poté již křivka standardně plynule klesá – např. *Plasmodium yoelii* na obr. 2
b) Křivka je **konvexní** (Maximum je na počátku, poté křivka více méně pravidelně klesá – např. *Bos taurus* a další na obr. 2.

Kromě těchto dvou výrazných typů může při kvalitativní analýze vylišit i určité proteomy, které mají výraznou tendenci fluktuovat – to je ovšem patrné pouze na jemnějším měřítku rozdělení slovníku na 10 částí, kde se projevuje obvykle křivka dvěma lokálními maximy (Např. *Paramecium t.* a *Tetrahymena t.* – obojí Ciliatea) – tato vlastnost je v kvalitativní analýze postižitelná, objevuje se ale jen u několika studovaných proteomů – všech patřících dle nového rozdělení eukaryot mezi Chromista.

Obr. 2: Distribuce unikátních amfipatických 12-gramů ve slovníku rozděleném na 3 části



Distribuční křivky frekvence unikátních amfipatických 12-gramů u 10 reprezentativních druhů. Průměrná křivka je označena přerušovanou čarou.

3. 2. Obecné vlastnosti slov u lidských jazyků

Jak lze předpokládat, v lidských jazycích dominuje alternace souhlásek a samohlásek, která je dána tím, že vyslovit lze fyzicky pouze určité kombinace hlásek a tudíž se v lidských jazycích nevyskytují dlouhé řetězce souhlásek nebo samohlásek. Tento trend je patrný u všech uvedených jazyků a u obou žánrů. Zjednodušení modelu tento obecný trend nemění.

4. Závěr

Smyslem této studie bylo ilustrovat analogii mezi alternací polárních a nepolárních aminokyselin v proteinech, která dává vzniknout specifické struktuře amfipatického alfa helixu a střídání souhlásek, a samohlásek v lidských jazycích. V obou případech zde působí přirozená omezení, která vedou ke vzniku informační struktury nesoucí nějaký význam (Eco 1976).

Distribuce amfipatických proteinů v peptidech se zdá být druhově specifická a distribuční křivka rozmístění unikátních amfipatických *n*-gramů by mohla sloužit jako tzv. *proteomic signature* (Osmanbeyoglu and Ganapathiraju 2011; Pietrokovski et al. 1990) Další výsledky, které nejsou uvedené přímo v článku, poukazují na určitá specifika distribuce amfipatických peptidů u některých parazitů, konkrétně u *Plasmodií*. Dle studie popisující strukturu T-receptoru vázajícího se na antigen (Berzofsky 1989) (Delisi and Berzofsky 1985), T-buňky imunitního systému preferují v 99% peptidy schopné v sekundární struktuře formovat amfipatický alfa helix. Celkový rozdíl mezi parazity a volně žijícími organismy ve využití amfipatických a slov pro relevantní délky 4-6 amokyselin (viz Zemková et al., 2016) prokázán nebyl, ale atypické chování distribuční křivky u některých skupin parazitů poukazuje, že specifické využití konkrétně amfipatických peptidů může hrát roli v úniku parazitů před hostitelem podobně jako celková redukce peptidického slovníku.

5. Literatura

Anděl J., (1998): Statistické metody. Matfyzpress, Praha

Berezovsky IN, Grosberg AY, Trifonov EN. 2000. Closed loops of nearly standard size: Common basic element of protein structure. Febs Letters. 466(2):283-286.

Berzofsky J. 1989. Structural features of t-cell recognition: Applications to vaccine design. Philosophical Transactions of the Royal Society of London B, Biological Sciences. 323(1217):535-544.

Delisi C, Berzofsky JA. 1985. T-cell antigenic sites tend to be amphipathic structures. Proceedings of the National Academy of Sciences. 82(20):7048-7052.

Dunker AK, Romero P, Obradovic Z, Garner EC, Brown CJ. 2000. Intrinsic protein disorder in complete genomes. Genome Informatics. 11:161-171.

Dunker AK, Silman I, Uversky VN, Sussman JL. 2008. Function and structure of inherently disordered proteins. Current opinion in structural biology. 18(6):756-764.

Eco U. 1976. Theory of semiotics. Bloomington: Indiana

University Press.

Lehmann WP. 1992. Historical linguistics: An introduction. London: Routledge Matisoff.

Likeš J., Laga J., (1978): Základní statistické tabulky. SNTL, Praha.

Maddieson I 2013. Consonant Inventories. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/1, Accessed on 2015-11-21.

Mant CT, Zhou NE, Hodges RS. 1993. Stabilizing peptide and proteinstructure. The amphipathic helix.39.

Osmanbeyoglu HU, Ganapathiraju MK. 2011. N-gram analysis of 970 microbial organisms reveals presence of biological language models. BMC bioinformatics. 12(1):12.

Pietrokovski S, Hirshon J, Trifonov E. 1990. Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. Journal of Biomolecular Structure and Dynamics. 7(6):1251-1268.

Popov O, Segal D, Trifonov EN. 1996. Linguistic complexity of protein sequences as compared to texts of human languages. Biosystems. 38(1):65-74.

Sheskin, David, J. 2011. Handbook of pararmetric and nonparametric statistical procedures. Chapmann and Hall.

Zemková M, Zahradník D, Mokrejš M, Flegr J. 2016. Reduction of peptide vocabulary in parasites – an evolutionary immune evasion strategy? (submitted)

Evolutionary landscape of human genome vocabulary

1. Úvod

Rozložením lidského genomu na sekvence o fixní délce dostaneme nukleotidový „slovník“ – respektive seznam všech takto dlouhých sekvencí, které se v genomu nachází. Seřadíme-li je dle frekvence výskytu, na předních místech seznamu se nacházejí téměř výhradně jednoduché repetitivně sekvence a rovněž tzv. Alu sekvence, což jsou nejčastější roztroušené repetitivně sekvence uvnitř lidského genomu (Obr 1 přílohy 3). Předpokládejme jednoduchý scénář evoluce genomu: Původní ancestrální sekvence byly (a jsou i v současnosti)jednoduché repetitivní sekvence schopné se rychle replikovat a poskytovaly tak materiál pro evoluci genomu. Všechna tato (co do frekvence) „excesivní slova“ lze tedy brát jako „generátory“ (*sequence generators*). Způsob, jak identifikovat ze seznamu excesivních slov právě generátory, spočívá v nalezení takových sekvencí, jejichž bodové mutace se objevují s nižší frekvencí. V článku jsou představeny seznamy potenciálních generátorů na příkladu lidského genomu pro délky slov 15 nukleotidů.

1. 1. Repetitivní DNA v genomu

Repetitivní DNA tvoří až dvě třetiny lidského genomu (de Koning et al. 2011), obecně se rozlišují dva typy, podle toho, jestli jsou jednotlivé opakující se motivy řazeny za sebou v blocích (*tandem repeats*), nebo jestli jsou rozptýleny uvnitř genomu na různých místech (*interspersed repeats*) Tandemové repetice se taky označují jako satelitní DNA, protože při izolaci DNA vytvoří na elektroforetogramu výrazně odlišný proužek, který má, díky repetitivnosti odlišný podíl bazí než ostatní DNA. Nacházejí se často v oblasti centromer, jediný satelit o známé funkci je alfa satelit, jež tvoří pravděpodobně funkční jádro centromery. Jiné funkce zatím nebyly detekovány a tak je tradičně repetitivní DNA považována za „odpad“ (*junk*).

Rozptýlené repetice, obecně transpozony a retrotranspozony, jsou schopné „skákat“ po genomu a množit se mechanismem typu copy-paste. Příkladem typickým pro primáty včetně člověka jsou alu sekvence, což je nejčastější typ rozptýlené repetitivní DNA v lidském

genomu, představuje totiž asi 10 % veškeré DNA člověka. Každý Alu element je sekvence o délce asi 350 nukleotidů, přičemž obsahují dvě oblasti o délce 130 bp, mezi nimiž je krátká AT-bohatá oblast. Vyskytují se v lidském genomu v počtu 1–1,5 milionu kopií (de Koning et al. 2011; Liu et al. 2009). Samy o sobě nemají vnitřní repetitivnost.

1. 2. Ancestrální sekvence a hledání generátorů

V předešlých Trifonovových pracech je nastíněn možný scénář vzniku genomu z jednoduchých repetitivních sekvencí vlivem expanze tripletů (Frenkel et al. 2013; Frenkel and Trifonov 2012; Trifonov and Bettecken 1997). Autoři předpokládají, že na počátku evoluce stály jednoduché nukleotidové triplety, jež měly schopnost se rychle replikovat. Expanze tripletů je známa z neurodegenerativních chorob a patrně souvisí s molekulárním mechanismem sklouznutí templátu během replikace DNA (Wells, 1996). Je spojena se skupinou tzv. agresivních tripletů GCC, CGA a GAA a jejich cirkulárních permutací. Všechny tyto permuatace se mohou objevit v lidském genomu ve formě tandemových repetic, a mají obecně schopnost snáze expandovat.

Hledáme tedy ancestrální sekvence repetitivní povahy. Předpokládejme, že na počátku evoluce genomu byla DNA tvořena pouze těmito repetitivními úseky a postupnými mutacemi, delecemi, inzercemi a dalšími přestavbami genomu se původní „text“ natolik rozrůznil, že jeho někdejší repetitivní povaha už není na první pohled viditelná. Nicméně i tak by měly v genomu zůstat nějaké stopy po repetitivnosti. Invazivní charakter tandemových repetic naznačuje, že právě mezi těmito elementy jsou „generátory“, od kterých se odvíjela evoluce a jednou z funkcí repetitivní DNA by tak bylo trvalé zajišťování genetického materiálu pro evoluci genomu (Frenkel and Trifonov 2012; Trifonov 2004). Dále mohou generátory zřejmě vznikat i transpozicí nebo multiplikací genomu. Obecně se vyskytují v neutrálních částech genomu a neinterferují do jiných aktivit v genomu. Jejich evoluce může vést k negativním, pozitivním či neutrálním změnám. V případě změny vedoucí k vylepšení nějaké funkce může být změna akumulována.

2. Metodika

Z lidského genomu byl pro urychlení výpočtu vybrán vzorek o velikosti 10MB, z tohoto souboru byly hledány potenciální generátory, které lze obecně nalézt dle následujícího algoritmu:

Vybereme libovolný 15-gram, nalezneme všechny jeho jednobodové mutace (kterých je potenciálně 3krát15 = 45) a zjistíme, která se v daném genomu vyskytuje nejčastěji. Od tohoto nejčetnějšího 15-gramu najdeme opět všechny jednobodové mutace a pokračujeme s nejčetnějším mutovaným řetězcem. Tento krok se opakuje do té doby, dokud lze nalézt nějakou mutaci četnější. Nejčetnější nalezená mutace se nazývá generátorem.
Tento algoritmus byl realizován pro ilustrační účely článku ve zjednodušené formě: Výchozím bodem pro hledání generátorů nebyly libovolné *n*- gramy, ale vybrané *n*-gramy s největší četností v seznamu všech 15-gramů, u kterých je předpokládáno, že skutečně generátory jsou. Od těchto sekvencí byly nalezeny všechny jednobodové mutace a spočítány jejich frekvence. Tento krok byl zopakován pro každou jednobodovou mutaci. Jestliže žádná z těchto jednobodových mutací není četnější než potenciální generátor, pak se jedná skutečně o generátor.

V seznamu 15-gramů byly identifikovány Alu sekvence na základě shodnosti s konsenzem pro Alu sekvence (Jurka and Milosavljevic 1991).

4. Výsledky a diskuse

Námi vybraná slova ze seznamu nejčetnějších slov (AAAAAAAAAAAAAAA, Alu sekvence TAATCCCAGCACTTTG, TGTGTGTGTGTGTGTG, TATATATATATATATA, GAGAGAGAGAGAGAGA, GCCGCCGCCGCCGCCG) se ukázaly být skutečně generátory (Obr. 1-4 přílohy 3) Zajímavým zjištěním je, že součet četností všech jejich jednobodových mutací je ve všech případech přibližně roven četnosti samotných generátorů. Rychlost mutací v obou směrech je tedy v rovnováze.

5. Závěr

Tato studie navrhuje, jak by mohlo být možné najít původní a současné generátory genomického slovníku. Neobvykle velký počet jednoduchých tandemových repetic v genomu naznačuje, že se jedná o generátory, neboť mechanismus jejich formace existuje od okamžiku jejich vzniku. Tyto sekvence lze tedy považovat za ancestrální sekvence původních genomů a zároveň za stavební materiál, který byl v následné evoluci využit.

Tento přístup by bylo možno dále rozvést na příkladech dalších organismů, a bylo by ho možné rozvinout např. i jako alternativní cestu při hledání fylogenetických vztahů.

5. Literatura

de Koning AJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS genetics. 7(12):e1002384.

Frenkel ZM, Barzily Z, Volkovich Z, Trifonov EN. 2013. Hidden ancient repeats in dna: Mapping and quantification. Gene. 528(2):282-287.

Frenkel ZM, Trifonov EN. 2012. Origin and evolution of genes and genomes. Crucial role of triplet expansions. J Biomol Struct Dyn. 30(2):201-210.

Jurka J, Milosavljevic A. 1991. Reconstruction and analysis of human alu genes. Journal of molecular evolution. 32(2):105-121.

Liu GE, Alkan C, Jiang L, Zhao S, Eichler EE. 2009. Comparative analysis of alu repeats in primate genomes. Genome research. 19(5):876-885.

Trifonov E, Bettecken T. 1997. Sequence fossils, triplet expansion, and reconstruction of earliest codons. Gene. 205(1):1-6.

Trifonov EN. 2004. Tuning function of tandemly repeating sequences: A molecular device for fast adaptation. Evolutionary theory and processes: Modern horizons. Springer. p. 115-138.

Wells RD. 1996. Molecular basis of genetic instability of triplet repeats. Journal of Biological Chemistry. 271(6):2875-2878.

Genome and language – two scripts of heredity (Ontogenetic theory of language origin)

"*If we were ever able to solve the enigma of glottogenesis, then we would possess a vital clue to the mystery of life itself."* (Marcel Danesi, 1993)

1. Úvod

V článku je analogizována otázka vzniku genomu a vzniku jazyka. Jak již bylo rozebráno v předešlém článku (Zahradník et al., 2015), evoluce genomu se pravděpodobně ubírala od jednoduchých tandemových repetic, jejichž invazivní charakter je dán expanzí tripletů založené na mechanismu sklouznutí templátu při replikaci (Frenkel and Trifonov, 2012; Trifonov and Bettecken, 1997; Wells, 1996). Frenkel a Trifonov (2012) ukazují, že původní, dnes mutacemi a jinými evolučními změnami překrytá repetitivní povaha genomu, je stále patrná a rekonstruovatelná z mRNA. Tento obecný, v celku intuitivní princip povstávání složitých struktur z jednoduchých repetitivních úseků DNA typu TGTGTG…, GCCGCCGCC…, schopných se množit není ničím novým a byl navržen například již v pracech Susumu Ohna (Ohno and Ohno, 1986; Ohno et al., 1968), nicméně nebyl dále systematicky rozvíjen. Podobně opomenutě si překvapivě stojí v lingvistice vznik lidského jazyka respektive řeči z prvních dětských repetitivních holofrází tzv. kanonického žvatlání (*Canonical babbling),* např. Ma-ma-ma-ba-ba-ba....

1. 1. Ontogeneze a fylogeneze jazyka

Přestože zájem o odhalení tajemného povstávání jazyka je prastarý (doložený už u Herodota [10]), lingvistika 20. století v podstatě toto téma zavrhla jako nevhodné pro akademickou půdu a exaktní zkoumání (Fitch 2005). Přestože práce druhé poloviny 19. Století se velmi intenzivně zaobíraly fenoménem dětského jazyka, prvních zvuků a jazykových patternů u dětí před osvojením vlastního mateřského jazyka[11]. Tyto myšlenkové přístupy k jazyku skrze dětskou řeč se objevují znovu až v 80. letech 20. Století (MacNeilage and Davis 1990; Vihman et al. 1986; Vihman et al. 1985). Studdert- Kennedy (1991) přímo

[10] Herodotos, kniha II, 4c-e, známý pokus faraona Psamtheka, který údajně odebral dvě děti matce a dal je vychovávat němému ovčákovi, aby zjistil, zda se u dětí objeví jazyk či nikoliv.
[11] Vyčerpávající přehled záznamů různých autorů týkajících se dětského jazyka a možnosti autonomního vzniku nových slov u dětí podává např. F. Čáda: Studium řeči dětské II. Vývoj dětské zásoby slovní. 1908, Dědictví Komenského, Praha.

navrhuje, že proces povstávání dětského jazyka je dobrým kandidátem pro studium paralel mezi ontogenezí a fylogenezí. Vzhledem k tomu, že řeč „nefosilizuje“ a nemůžeme téměř nijak dokumentovat její vznik[12] (může nám být vývoj dětského jazyka vodítkem k odhalení evoluce jazyka. Bariérami k tomuto přístupu se však zdá být i zavedená lingvistická a pedagogicko-psychologická terminologie, která například dělí fáze jazyka na „pre-lingvistickou“ a fázi osvojování jazyka. Díky tomu období před počátkem „prvních slov“ nebylo bráno dlouho v potaz a fáze jako žvatlání (*baby babbling*) se považovalo za nerevlevantní pro vznik jazyka, který je přeci osvojován napodobováním a přijímáním již existujícího jazyka z okolního prostředí dítěte. Současné studie ale ukazují, že vývoj řeči je kontinuální proces začínající již v prenatálním stádiu.

1. 2. Existuje univerzální stádium jazyků?

Má-li být ontogeneticko – fylogenetický přístup k vývoji jazyka respektive řeči platný, pak by mělo být možné najít nějaké stadium v jejich vývoji, které je sdíleno všemi ostatními jazyky. Domníváme se, že tímto stadiem by mohla být právě perioda kanonického žvatlání, během kterého jsou dětmi generovány jednoduché zvukové sekvence repetitivního charakteru. Tyto repetitivní sekvence jsou zároveň široce sdílené v prvních dětských slovech označujících rodinu, základní potřeby a emoce atd. napříč světovými jazyky. Samotná perioda kanonického žvatlání začíná obvykle 7. měsícem a okolo 10. měsíce života dítěte často plynule přechází (i s delším časovým překryvem až několika měsíců) do fáze prvních „slov“. Zde je třeba poznamenat, že slovo nemusí být vždy nutně totožné s existujícími slovy mateřského jazyka, ale obecně je to „něco, co nese význam“. Zatímco kanonické žvatlání je (alespoň ve svých počátečních fázích) čistě reflexivní činnost, při které dítě trénuje čelisti a patřičné svaly, přechod ke slovům je charakterizován spojením zvuku a významu. Vzhledem k tomu, že u všech dětí jsou reflexivně produkovány podobné zvukové sekvence (protože jejich podoba je dána přirozenými constraints vyvíjejícího se jazykového aparátu) (MacNeilage and Davis 1990; Olson 2003) není překvapením, že první slova jsou sdílená napříč jazyky a pocházejí nejčastěji právě z repertoáru sekvencí kanonického žvatlání. Rodiče samozřejmě pozitivně reagují na první hlasové projevy svého potomka, které, na rozdíl od broukání již připomínají, byť zdánlivě, řeč, a přikládají jim význam - např. označení nejbližších osob apod. Roli zde

[12] Pokusy o hledání indicií ve fosilním materiálu zde nicméně jsou, např srovnávání postavení čelistí a lebky u neandrtálců a moderního člověka, viz Lieberman P. 1984. The biology and evolution of language. Harvard University Press..

pak má i hravost dítěte – pocity libosti či jiných emocí při vyslovování některých zvuků. Ostatně důležitou roli repetitivnosti ve vyjádření pocitů pozorujeme i v běžné řeči (opakování důležitého) nebo v hudbě (existence refrénů, opakování motivů atd.) Postupně tak přechází žvatlací reflexivní fáze ve fázi vědomou, kdy jsou jednoduché repetitivní holofráze samotným dítětem ztotožňovány s určitými významy.

Důležitost periody kanonického žvatlání dokazují i některé lékařské studie: Když dítěti v příslušné fázi bylo zabráněno z důvodu tracheostomie v produkci patřičných zvuků, plné řečové schopnosti pak u něj již nebylo nikdy zcela dosaženo (Locke and Pearson, 1990). Nicméně i v tomto případě byly pozorovány po ukončení tracheostomie alespoň jednoduché náznaky typických patternů kanonického žvatlání. Newport a Meier (1985) referují případ, kdy hluchoněmé děti vystavené od narození znakovému jazyku „žvatlaly“ pomocí prstíků. Přestože se tvrdilo, že hluchoněmé děti obecně vynechávají periodu kanonického žvatlání, což bylo bráno jako argument proti jeho univerzalitě, Lynch a kol. (1989) dokládají, že ani úplná hluchota nemusí nutně zabránit produkci kanonického žvatlání, pokud je patřičně stimulováno.

Univerzálnost zvukových patternů během kanonického žvatlání a sdílení napříč jazyky prokazují četné studie např. Gildersleeve-Neumann et al., 2013; Lee et al., 2010; Teixeira and Davis, 2002.

2. Metodika

Z webových zdrojů a z osobních interview autorů byl sestaven mezinárodní „slovník“ dětských repetitivních slov. Byly brány pouze slova vycházející z jednoduchých repetic typu ma-ma, da-da, te-te atd. a k nim přiřazovány významy v jednotlivých jazycích. Byla vybírána pouze slova základní užitá pro nejbližší okolí a emoce nikoliv například modifikace poukazující spíše na druhotný vznik z již existujícího slova užívaného v jazyce dospělých. Nebyly zahrnuty ani jednoduché „mutace“ – odvozeniny typu nany, papu, hají apod. (Tabulka 1,2 přílohy 4)

3. Výsledky a diskuse

Předložený slovník dětských repetitivních slov spolu s již uvedenými četnými studiemi poukazují na to, že perioda kanonického žvatlání je univerzálním stádiem lidské řeči sdíleným napříč jazyky. Tato skutečnost nás rovněž opravňuje k tvrzení, že z ontogeneze dětské řeči můžeme rekonstruovat alespoň částečně evoluci lidské řeči obecně.

Problematičnost tohoto tématu v rámci současné lingvistiky vidíme mimo jiné ve striktním oddělování řeči a jazyka. Řeč jako individuální projev buď není vůbec brána v potaz a důraz je dáván pouze na jazyk jako abstraktní systém, nebo jsou oba fenomény vágně zaměňovány. Současná biolingvistika, která se snaží o nahlédnutí jazyka jakožto biologického fenoménu, bohužel hledá pouze doklady biologického zakořenění jazyka v konkrétních genech. Předpokládá, dle Chomského (1986), že jazyk je invariantou vzniklou například nějakou náhodnou mutací, apriorním vnitřním schématem daným exklusivně pouze lidskému druhu. Tento pohled bohužel uzavírá možnosti evolučního přístupu k jazyku. Z hlediska historie vědy zde ostatně můžeme vidět zajímavou paralelu mezi neodarwinistickou doktrínou demonstrovanou například Monodem (1971), která zachází s živými bytostmi jako s pasivními objekty danými invariantně pouze svými geny. Podobně byl jazyk v moderní lingvistice odtržen od svého zdroje – řeči a ztotožněn s matematicky popsatelným abstraktním schématem, které je dáno apriorně.

Závěr

Analogie mezi lidskými jazyky (objekty lingvistiky) a živými bytostmi (objekty přírodních věd) nespočívá v porovnávání sekvencí jazyka či genů, nebo matematicky definovaných charakteristik linearizovaných genetických zápisů s lidskými texty. Stejně tak zde není na místě otázka, zda je jazyk nějak zapsán v genetické informaci. Naopak tato analogie připomíná, že ani jazyk ani život nejsou pouze lineární invariantní zápisy, ale naopak živoucími autonomními a sebeutvářejícími se strukturami. Otázka po vzniku života je tudíž podobně zastřené povahy jako otázka po vzniku jazyka.

Jak v případě evoluce genetických sekvencí, tak i u lidského jazyka pozorujeme povstávání komplexních struktur nesoucích významy z jednoduchých repetitivních motivů, které následně mutují a diverzifikují.

Literatura

Chomsky N. 1986. Knowledge of language: Its nature, origin, and use. Greenwood Publishing Group.

Danesi M. 1993. Vico, metaphor, and the origin of language. Indiana University Press.

de Koning AJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS genetics. 7(12):e1002384.

Čáda F. 1908. Studium řeči dětské II. Vývoj dětské zásoby slovní. Dědictví Komenského, Praha.

Fitch WT. 2005. The evolution of language: A comparative review. Biology and philosophy. 20(2-3):193-203.

Frenkel ZM, Barzily Z, Volkovich Z, Trifonov EN. 2013. Hidden ancient repeats in dna: Mapping and quantification. Gene. 528(2):282-287.

Frenkel ZM, Trifonov EN. 2012. Origin and evolution of genes and genomes. Crucial role of triplet expansions. J Biomol Struct Dyn. 30(2):201-210.

Gildersleeve-Neumann CE, Davis BL, MaCNeilage PF. 2013. Syllabic patterns in the early vocalizations of quichua children. Applied Psycholinguistics. 34(01):111-134.

Lee SAS, Davis B, MaCNeilage P. 2010. Universal production patterns and ambient language influences in babbling: A cross-linguistic study of korean- and english-learning infants*. Journal of Child Language. 37(02):293-318.

Lehmann wP. 1992. Historical linguistics: An introduction. London: Routledge Matisoff.

Lieberman P. 1984. The biology and evolution of language. Harvard University Press.

Locke JL, Pearson DM. 1990. Linguistic significance of babbling: Evidence from a tracheostomized infant. Journal of Child Language. 17(01):1-16.

Lynch MP, Oller DK, Steffens M. 1989. Development of speech-like vocalizations in a child with congenital absence of cochleas: The case of total deafness. Applied Psycholinguistics. 10(03):315-333.

MacNeilage PF, Davis B. 1990. Acquisition of speech production: Frames, then content.

Monod J. 1971. Chance and necessity. New York: Alfred A.Knopf.

Newport EL, Meier RP. 1985. The acquisition of american sign language. Lawrence Erlbaum Associates, Inc.

Ohno S, Ohno M. 1986. The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition. Immunogenetics. 24(2):71-78.

Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. Hereditas. 59(1):169-187.

Olson S. 2003. Mapping human history: Unravelling the mystery of adam and eve. Bloomsbury.

Studdert-Kennedy M. 1991. Language development from an evolutionary perspective. Biological and behavioral determinants of language development.5-28.

Teixeira ER, Davis BL. 2002. Early sound patterns in the speech of two brazilian portuguese speakers. Language and Speech. 45(2):179-204.

Trifonov E, Bettecken T. 1997. Sequence fossils, triplet expansion, and reconstruction of earliest codons. Gene. 205(1):1-6.

Trifonov EN. 2004. Tuning function of tandemly repeating sequences: A molecular device for fast adaptation. Evolutionary theory and processes: Modern horizons. Springer. p. 115-138.

Vihman MM, Ferguson CA, Elbert M. 1986. Phonological development from babbling to speech: Common tendencies and individual differences. Applied Psycholinguistics. 7(01):3-40.

Vihman MM, Macken MA, Miller R, Simmons H, Miller J. 1985. From babbling to speech: A re-assessment of the continuity issue. Language.397-445.

Wells RD. 1996. Molecular basis of genetic instability of triplet repeats. Journal of Biological Chemistry. 271(6):2875-2878.

Přílohy:

1. Zemková M., Zahradník D, Mokrejš M, Flegr J. (2016) Reduction of peptide vocabulary in parasites – an evolutionary immune evasion strategy? (Submitted)

Abstract:

Self-nonself discrimination by vertebrate immune systems is based on the recognition of the presence of peptides in proteins of a parasite that are not contained in the proteins of a host. Therefore, a reduction of the number of 'words' in its own peptide vocabulary could be an efficient evolutionary strategy of parasites for escaping recognition. Here, we compared vocabularies of 30 endoparasitic and 17 free-living unicellular organisms and also 8 multicellular parasitic and 16 multicellular free-living organisms. We found that both unicellular and multicellular parasites used a significantly lower number of different pentapeptides than free-living controls. Impoverished pentapeptide vocabularies in parasite was observed in all five clades that contains both the parasitic and free-living species. The effect of parasitism on number of peptides used in an organism's proteins is larger than all other studied factors, including the size of a proteome, the number of coded proteins, etc. This decrease of peptide diversity was partly compensated for by an increased number of hexapeptides. Our results support the hypothesis of parasitism-associated reduction of peptide vocabulary and suggest that T-cell receptors mostly recognize the five amino acids long part of peptides that are presented in the groove of MHC molecules.

One Sentence Summary:

Parasites differ from free- living organisms in their peptide vocabulary usage, e.g. in the number of different peptides of defined length used in proteome. Parasites use impoverished pentapeptide and enriched hexapeptide vocabularies, which could protect them against recognition by vertebrate-MHC-based host immune system. Our results refer to a critical role of peptide length in the process of T-cell recognition.

Self-nonself discrimination in vertebrates is based on detection of peptides in proteins of parasites which are not present in proteins of a host. Therefore, parasites are under a strong selection pressure to eliminate maximum possible number of peptides – the potential targets for the hosts' immunity – from their proteomes. We found that both unicellular and

multicellular parasites use impoverished pentapeptide vocabulary in comparison with free-living organisms. The effect of parasitism on number of peptides used in an organism's proteins is larger than all other studied factors, including the size of a proteome and the number of coded proteins. Our results bring a new understanding to the phenomenon of parasitic molecular mimicry and our findings could be potentially applied in vaccine design.

Introduction

The immune system recognizes the presence of proteins of foreign origin by the occurrence of peptides that are not present in a host's own proteins.

As a part of the host-parasite evolutionary arms race, a parasite could decrease the probability of its recognition by reducing the number of different peptides ('words') in its vocabulary (vocabulary reduction), and by mimetizing the peptide vocabulary of its host (vocabulary mimicry). It could therefore be expected that parasitic organisms will have lower number of different peptides in their proteome than free-living organisms.

Almost all cells in vertebrate bodies (except e.g. for sperm and trophoblast) (*1*) present fragments of proteins, i.e. short peptides, on their surface (*2*). These peptides are captured in the grooves of MHC class I molecules on the surface of somatic cells and MHC class II molecules on the surface of specialized antigen presenting cells (APC). The peptides can be recognized as non-self by the T-cells that carry molecules of a T-cell receptor with matching specificity. Each young T-cell carries one type of T-cell receptor with an affinity toward certain self or non-self peptide. The population of T-cells is subjected to negative selection in a thymus. In this process, all T-cells carrying a receptor that recognizes any peptide presented in a thymus die or are functionally incapacitated. Therefore, when a mature T-cell recognizes a peptide outside the thymus, it is most probably a non-self peptide that originated from the proteins of a parasitic organism. Such proteins are either synthesized in a particular host cell (typically the peptides of viral origin presented on MHC class I molecule) or originated from the proteins of parasites engulfed by APC (typically the peptides presented on MHC class II molecule) (*3- 5*).

The mechanism of MHC-based self/non-self discrimination suggests that the number of different peptides in a vocabulary, i.e. the number of potential objects for T-cell recognition, is a critical parameter in the host-parasite arms race and therefore also an

important target of natural selection in parasitic organisms. The standardized size of any vocabulary can be expressed as the vocabulary usage, the ratio of the actual vocabulary size (the number of all different words) to the maximal combinatorially possible vocabulary size (*6,7*). (For details see equation 1 and 2 in supplementary material.)

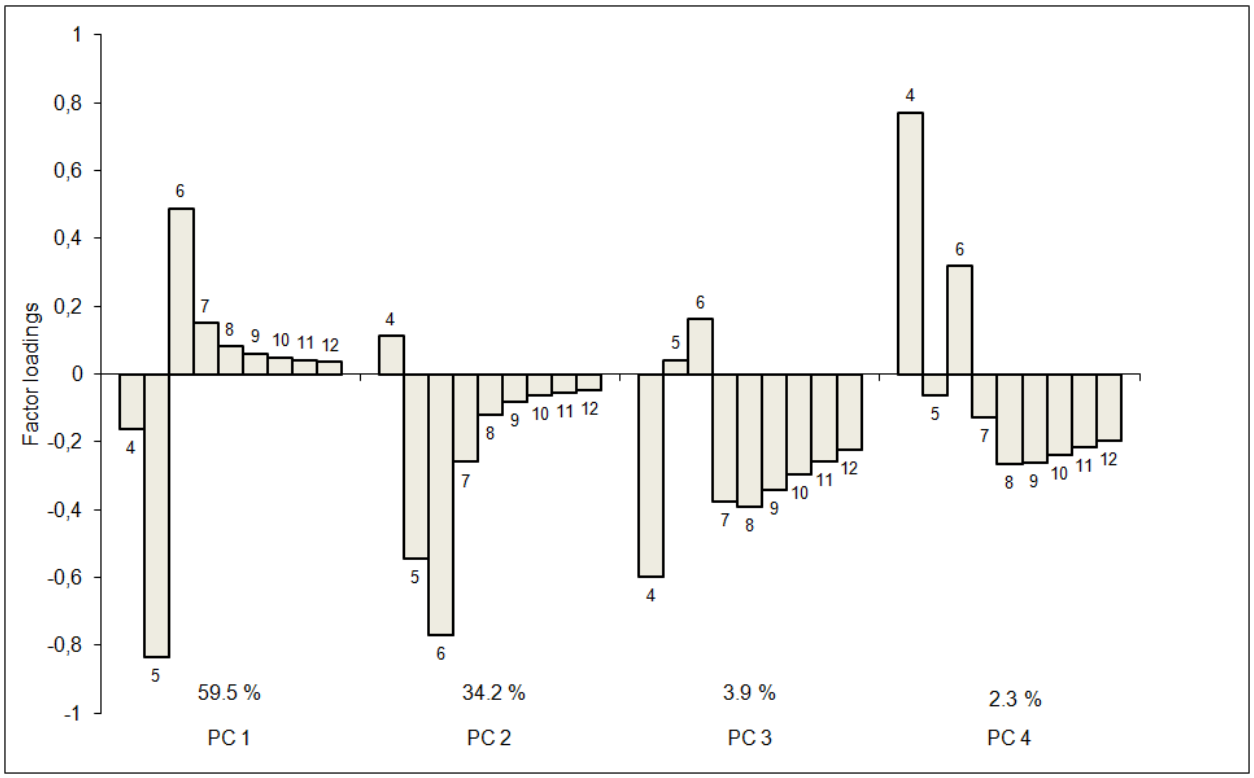
Here we searched for indices of reduced peptide vocabulary in parasites by comparing the vocabulary usage of proteomes of 38 endoparasites (8 multicellular parasitic helminths, 30 unicellular protozoan parasites) with 33 free-living eukaryotic organisms.

Results

Vocabulary usages for peptides of lengths from 4 to 12 amino acids (U_4–U_{12}) were computed for 71 proteomes of different organisms. (Materials and methods with the List of organism S1 are available as supplementary materials on *Science* Online.) Individual values of U_4–U_{12} of all proteomes are listed in the Supplementary table S3.

As nine variables U_n were highly correlated, we used the method of principal component analysis (PCA) to reduce the number of variables and to obtain independent composite variables – the principle components (PC). The first four PC had eigenvalues higher than 1 and explained 99.9% of the variability in vocabulary usage (figure 1). The first two principal components (PC1 and PC2) were loaded mostly with pentapeptide and hexapeptide vocabularies (U_5 and U_6). PC1 was negatively loaded with pentapeptides and positively with hexapepeptides, while PC2 was negatively loaded with by all types of peptides, except tetrapeptides. PC3 was positively loaded with hexapeptides and negatively by tetrapetides, and peptides longer than seven amino acids (U_7–U_{12}). PC4 was positively loaded primarily by tetrapetides, and also partly by hexapeptides.

Figure 1: Four principal components explain 99.9 % of interspecies variability in peptide vocabulary usage.



The figure shows particular factor loadings and corresponding percentages of explained interspecies variability in vocabulary usage. Column labels 4-12 indicate the length of the peptides, which load the particular principle component.

Vocabulary usage can be influenced by parasitism and also by various non-ecological factors, such as the complexity of an organism, genome redundancy, etc. Therefore, we used simple multivariate ANCOVAs to find out which parts of interspecies variability in vocabulary usage (independent variables PC1-PC4) could be explained by four factors reflecting the size and redundancy of proteomes (size of proteome, size of non-redundant part of proteome, count of proteins in whole proteome and count of proteins in non-redundant part of proteome) and which parts could be explained by parasitism or other binary factors. Because of the nested character of the data, a separate multivariate analysis was performed for each binary variable of interest, namely for parasitism (parasites vs. free-living organisms), unicellular parasitism (unicellular parasites vs. unicellular free-living organisms), multicellular parasitism (multicellular parasites vs. multicellular free-living organisms), multicellularity (unicellular vs. multicellular organisms), endoparasitism (extracellular parasites vs. intracellular parasites) and heterotrophy (heterotrophs vs. autotrophs), see Table 1 and Supplementary table S1.

Table 1: Effects of parasitism, form of parasitism, multicellularity, and heterotrophy on peptide vocabulary usage.

	PC1 (59.5%) $\downarrow U_5 \uparrow U_6$			PC2(34.2%) $\downarrow U_5 \downarrow U_6$			PC3 (3.9%) $\downarrow U_{4,7\text{-}12} \uparrow U_6$			PC4 (2.3%) $\uparrow U_4, \uparrow U_6, \downarrow U_{7\text{-}12}$		
	beta	%	p-value	beta	%	p-value	beta	%	p-value	beta	%	p-value
Parasitism– Non-parasitism	0.0661	6.5	**0.000**	0.0332	2.8	0.137	-0.0158	5.7	**0.041**	0.0077	2.3	0.196
Unicellular parasitism.– Unicell. non-parasitism	0.0567	5.7	**0.007**	0.0345	2.4	0.220	-0.0038	0.3	0.680	0.0095	2.7	0.260
Multicell. par. – Multicell. non-parasitism	0.0627	5.3	**0.001**	-0.0190	1.6	0.545	-0.0296	16.4	0.057	0.0081	6.2	0.190
Unicellularity – Multicellularity	0.0409	2.6	**0.008**	0.0761	15.6	**0.000**	-0.0012	0.0	0.877	-0.0108	4.6	0.064
Unicell. parasitism – Multicell. parasitism	0.0481	6.5	0.105	0.1237	17.7	**0.002**	0.0044	0.2	0.759	0.0029	0.1	0.825
Unicell. non-parasitism– Multicell. non-parasitism	0.0412	3.8	**0.025**	0.0709	14.5	**0.022**	-0.0076	1.5	0.493	-0.0125	9.5	0.076
Intracelullar parasitism– Extracellular parasitism	-0.0121	0.7	0.585	0.0207	0.8	0.495	0.0115	2.2	0.275	-0.0282	17.4	**0.004**
Heterotrophy – Autotrophy	0.0217	0.7	0.189	-0.0374	3.4	0.103	-0.0092	1.8	0.255	0.0165	9.8	**0.006**

The table summarizes the results of analyses of 32 simple multivariate ANCOVA models with five independent variables: size of proteome, size of non-redundant part of proteome, number of proteins in proteome, number of proteins in non-redundant part of proteome and one of focal binary variables listed in the first column. The columns 2-4, 5-7, 8-10 and 11-13 show results (slope beta, % of explained variability, and significance of two-sided test) for four dependent variables, namely (PC1-4). Significant p-values are printed in bold. Positive beta value means that the first group of organisms of the compared pair has a higher particular PC_n value than the second group of organisms. For example in the first row, parasites have significantly higher PC1 values than free-living organisms. Signs of correlation of PC1-4 with vocabulary usage are indicated with an arrow in the legend of each principle component and the size of the arrow illustrates the strengths of this effect, for details see Fig. 1.

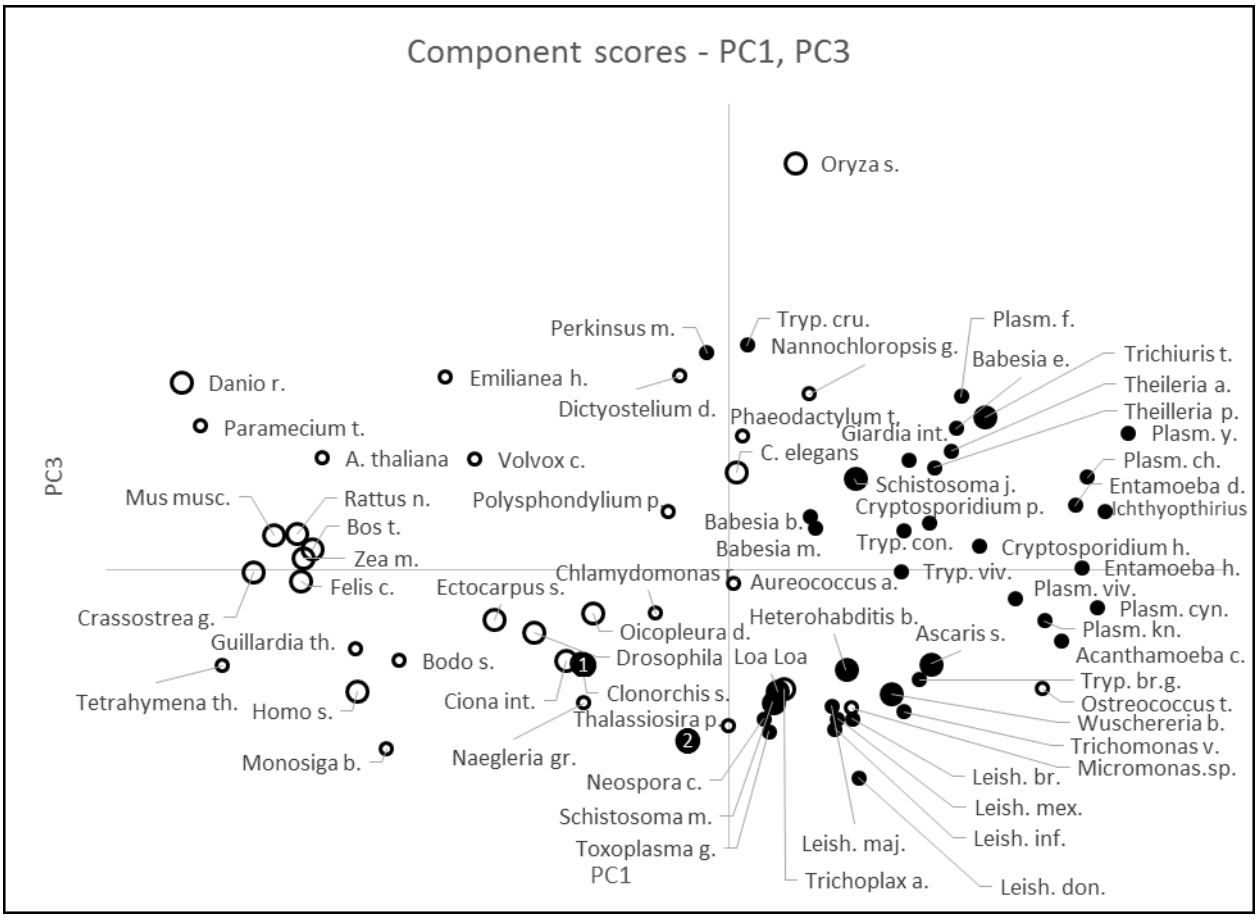
The strongest and the most significant effect observed in our data was a positive effect of parasitism on the PC1, i.e. on the variable explaining the largest part of interspecies variability in the vocabulary usage. Parasites had relatively impoverished pentapeptides and tetrapeptides, and relatively enriched hexapeptide vocabularies. Parasitism explained nearly 6.5 % of this variability, while, for example, all four variables describing size and redundancy of proteome together explained less than 4 % of this variability (Supplementary table S4). The main factor influencing PC2 was unicellularity/multicellularity of organisms. Unicellular organisms had relatively impoverished vocabularies (especially the pentapeptides and hexapeptides vocabularies), except the tetrapeptides vocabulary. PC3 was negatively influenced by parasitism; the parasites had relatively impoverished pentapeptides and hexapeptides vocabularies and enriched all other vocabularies. Extracellular parasites (both unicellular and multicellular) had generally higher values of PC4, i.e., they had relatively enriched tetrapeptide and hexapeptide vocabularies and relatively impoverished other vocabularies (in comparison with intracellular parasites, not with free living organisms).

The distribution of parasitic and free-living organisms within the two-dimensional space as defined by two principal components correlated with parasitism (PC1 and PC3) is shown in Figure 2. Parasitic organisms are clustered on the right side of the graph. Multicellular free-living and parasitic organisms are shifted left-down in comparison with unicellular free-living and parasitic organisms. The two-dimensional space defined by PC1

(correlated with parasitism) and PC2 (correlated with multicellularity) is shown in Supplementary figure S1 and the two-dimensional space defined by PC1 (correlated with parasitism) and PC4 (correlated with intracellularity) is shown in Supplementary figure S2.

Because of the existence of a phylogenetic relation between the analyzed organisms, the results of our statistical analysis could be influenced by the effect of pseudoreplications. To eliminate this effect, we used an exact pair test. Only five clades contained both parasitic and free-living organisms (Kinetoplastids, Ciliates, Nematodes, Opisthokonts and Sar). Within all five pairs, the mean value of PC1 was higher for the parasitic than for the free-living organisms ($p = 0.031$, exact variant of binomial sign test for dependent samples (*8*)) No correlation existed between PC1 and PC3. Therefore, it was possible to also compare the mean PC3 of parasitic and free-living organisms within the same 5 clades. This time, the mean value of PC3 was lower for parasitic organisms in all 5 clades. Global tests for all ten pairs (Supplementary table S5) showed highly significant ($p < 0.001$) support for our hypothesis of lower peptide vocabulary diversity (namely higher PC1 and lower PC3) in parasitic organisms.

Figure 2: Effect of parasitism on peptide vocabulary usage



Two-dimensional space defined by two principle components correlated with parasitism (PC1 and PC3). Dark and white circles denote the position of parasitic and free-living organisms, respectively. Larger circles indicate multicellular organisms. We used two proteome datasets for parasitic trematode Clonorchis sinensis – number 1 corresponds to a proteome derived from a set of genome-based proteins (assembly v2.0 from NCBI Genomes) whereas the number 2 corresponds to a transcriptome-based predicted proteome (from HelmDB).

Discussion

Our study suggests that the number of different peptides per proteome could reflect an ecological strategy of this species, namely the difference between parasitic and non-parasitic organisms. Nearly all variability (99.9%) of vocabulary usage in eukaryotic organisms was explained by four principle components. PC1, the factor explaining nearly 60 % of interspecies variability (59.5 %), was influenced most strongly by the effect of parasitism and less strongly by the length of a filtered proteome, i.e. size of non-redundant part of the proteome. PC2 was influenced by the unicellularity/multicellularity of the organism, PC3 was influenced by parasitism in multicellular organisms and PC4 reflected differences between extracellular and intracellular parasites.

From all factors studied, including the lengths of proteomes, parasitism had the strongest effect on peptide vocabulary usage. We detected this effect independently in two

sets of organisms, multicellular organisms and unicellular organisms. The results suggest that parasites have lower diversity of pentapeptides, which is partly compensated for by higher diversity of hexapeptides. It can be hypothesized that T-cells recognize peptides of five aminoacid residues in length when attached in the groove of MHC I protein. The length of trimmed peptides, which are loaded onto MHC I protein, is about 8 – 10 amino acids and the length of those loaded on MHC II is even higher. However, it was experimentally shown that only the residues at the top of the binding groove are recognized by T-cell receptors while those at the bottom of the groove are used to bind the peptide to the groove of MHC protein (*9*). Peptides usually contain only 2 or 3 amino acids that are critical for T-cell recognition; however, to trigger the response of the T-cell receptor the peptides must be longer by at least one or two additional amino acid residues (*10*). This agrees with our observation that parasites have the most strongly impoverished pentapeptide and partly impoverished tetrapeptide vocabulary. It is highly probable that for the preservation of functionality of proteins some minimal 'linguistic' complexity is required. Therefore, reduction at the level of pentapeptides and tetrapeptides should probably be compensated for by a richer hexapeptide vocabulary.

The length of actually recognized parts of peptides on MHC I is lower than on MHC II (*11*). The MHC I and MHC II present mostly peptides from intracellular and extracellular parasites, respectively. Therefore, we could expect that the type of parasitism (intracellular vs. extracellular) should affect the vocabulary usage – the intracellular parasites should have more impoverished shorter peptides-vocabularies than longer peptides -vocabularies. Indeed, we observed the effect of this type of parasitism on PC4, i.e. on the factor loaded mostly by high values of tetrapetides. The effect of intracellularity/extracellularity on vocabulary usage was relatively weak. It must be noted, however, that some extracellular antigens can also be presented through the MHC I pathway (in a process known as cross-presentation) (*5, 11, 12*) and that many seemingly intracellular parasites (such as representatives of the phylum Apicomplexa) in fact occupy an interior of parasitophorous vacuole – the organelle in some respects homologous to an extracellular, rather than intracellular, compartment (*13*).

Differences between parasites and free-living organisms are clearly visible from Figures of component scores where parasites are aggregated in a region of positive PC1 values while free-living organisms are clustered in the region of negative PC1 values. There are some interesting exceptions to this trend. A rich peptide vocabulary (including pentapeptides) of *Perkinsus marinus* can be explained by the fact that its host oyster does not possess an MHC-based immune system. Although free-living, the *Ostreococcus tauri* has a

highly reduced, parasite-like, peptide vocabulary. This tiny green alga is the smallest and the most reduced autotrophic eukaryote in our dataset so the reduction of its vocabulary could be related to its extreme simplicity. We have no explanation for the non-parasite-like (overly rich) vocabulary of the trematode *Clonorchis sinensis*, except possible undetected contamination by genes from cat liver tissue or from any sample processed in the respective sequencing lab (*14*). When purely transcriptome-based predicted proteins from Young et al. (*15)* were analyzed (downloaded from HelmDB), the position of *C. sinensis* in the two dimensional space of PC1 and PC3 moved towards the cluster of parasitic organisms (Figure 2).

Though it was not a subject of the present study, we detected the effect of multicellularity on the second strongest principal component. Multicellular organisms, both parasitic and free-living, have relatively richer hexapeptide and pentapeptide vocabularies, which could be an effect of the higher complexity of multicellular organisms. It must be noted, however, that only representatives of three phyla of multicellular organisms (Metazoa, Metaphyta and Charophyta) were included in this analysis. Therefore, this result may be biased by the effect of pseudoreplications.

Four independent lines of evidence, namely impoverished vocabulary in unicellular parasites, multicellular parasites, results of a phylogenetic contrast test performed on 5 pairs of sister taxa, and the fact that *Perkinsus* (one of the two parasites of hosts lacking MHC in the analyzed dataset) has unreduced peptide dictionary are in an agreement with our *a priori* hypothesis about the reduced peptide vocabulary of parasitic organisms. Other explanations of the observed pattern, for example the theoretical possibility of the reduction of non-housekeeping proteins in parasites, are of course also legitimate and should be tested when necessary proteomes become available. The results also suggest that T-cells recognize MHC-attached peptides of lengths 4-5 amino acids, which could possibly be of importance in vaccine construction. Our analysis included all proteomes larger than 1.2MB, which were available by May 2015. It would be possible to reproduce our findings with future, newly appearing proteomes as additional independent datasets. Similarly, it will be possible to use the developed software for testing the related peptide vocabulary mimicry hypothesis by studying similarities of peptide vocabularies between parasites and their specific hosts.

References and Notes:

1. R. C. King, W. D. Stansfield, P. K. Mulligan, *A dictionary of genetics* (Oxford Univ. Press, New York, 2006).

2. A. Lanzavecchia, Antigen-specific interaction between T-cells and B-cells. *Nature* **314**, 537-539. (1985). DOI:10.1038/314537a0

3. A.Craiu, T. Akopian, A. Goldberg, K. L. Rock, Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. *Proc. Natl. Acad. Sci.* **94**, 10850-10855. (1997). DOI: 10.1073/pnas.94.20.10850

4. J. Neefjes, H. Ovaa, A peptide's perspective on antigen presentation to the immune system. *Nat Chem Biol* **9**, 769-775. (2013). DOI: 10.1038/nchembio.1391

5. E. S.Trombetta, I. Mellman, Cell biology of antigen processing in vitro and in vivo. Annu. Rev. Immunol. **23**, 975-1028. (2005). DOI: 10.1146/annurev.immunol.22.012703.104538

6. Y. L. Orlov, V. N. Potapov, Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic acids research* **32**, 628-633. (2004). DOI: 10.1093/nar/gkh466

7. O. Popov, D. Segal, E. N. Trifonov, Linguistic complexity of protein sequences as compared to texts of human languages. Biosystems **38**, p. 65-74. (1996). DOI:10.1016/0303-2647(95)01568-X

8. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical procedures* (Chapmann & Hall, ed.5, 2011).

9. A. Townsend et al., The epitopes of influenza nucleoprotein recognized by cytotoxic T lymphocytes can be defined with short synthetic peptides. Cell **44**, 959-968. (1986). DOI:10.1016/0092-8674(86)90019-X

10. J. M. Austyn, K. J. Wood, *Principles of Cellular and Molecular Immunology* (Oxford Univ. Press, New York, 1994), pp. 148-161.

11. J. M. Vyas, A. G. Van der Veen, H. L. Ploegh, The known unknowns of antigen processing and presentation. *Nature Reviews Immunology* **8**, 607-618. (2008). DOI: 10.1038/nri2368

12. P. Paz, N. Brouwenstijn, R. Perry, N. Shastri, Discrete Proteolytic Intermediates in the MHC Class I Antigen Processing Pathway and MHC I–Dependent Peptide Trimming in the ER. *Immunity* **11**, 241-251. (1999). DOI: 10.1016/S1074-7613(00)80099-0

13. K. Lingelbach, K. A. Joiner, The parasitophorous vacuole membrane surrounding Plasmodium and Toxoplasma: an unusual compartment in infected cells. *J Cell Sci* **111**, 1467-75. (1998).

14. X. Wang et al., The draft genome of the carcinogenic human liver fluke Clonorchis sinensis. *Genome Biol* **12**,p R107. (2011). DOI: 10.1186/gb-2011-12-10-r107

15. N. D. Young et al., Unlocking the transcriptomes of two carcinogenic parasites, Clonorchis sinensis and Opisthorchis viverrini. *PLoS Negl Trop Dis* **4**, p. e719. (2010) DOI: 10.1371/journal.pntd.0000719

16. M. Zemková, E. N. Trifonov, D. Zahradník, One common structural feature of "words" in protein sequences and human texts. *J Biomol Struct Dyn* **32**, 1085-91. (2014). DOI: 10.1080/07391102.2013.809317

17. A. C. Rencher, *Methods of multivariate analysis* (Wiley, New York, 2002), pp. 380-408.

18. K. Tyler, D. Engman, The life cycle of Trypanosoma cruzi revisited. *International journal for parasitology* **31**, 472-481. (2001). DOI: 10.1016/S0020-7519(01)00153-9

19. S. M Adl et al., The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology* **59**, 429-514 (2012). DOI: 10.1111/j.1550-7408.2012.00644.x

20. L. S. Diamond, C. G. Clark, A Redescription of Entamoeba Histolytica Schaudinn, 1903 (Emended Walker, 1911) Separating It From Entamoeba Dispar Brumpt, 19251. *Journal of Eukaryotic Microbiology* **40**, 340-344. (1993). DOI: 10.1111/j.1550-7408.1993.tb04926.x

21. Z. S. Hamzah, Petmitr M. Mungthin, S. Leelayoova, P. Chavalitshewinkoon-Petmitr, Differential detection of Entamoeba histolytica, Entamoeba dispar, and Entamoeba moshkovskii by a single-round PCR assay. *Journal of Clinical Microbiology* **44**, 3196-3200. (2006). DOI: 10.1128/JCM.00778-06

22. A. M Elliott, *Biology of Tetrahymena* (Hutchinson & Ross., Dowden, 1973)

Acknowledgments:

For all statistical computations, standard Base-package of R software was used (R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.) URL http://www.R-project.org/.

All data reported in this paper are available at NCBI: http://www.ncbi.nlm.nih.gov/ and Sanger institute: https://www.sanger.ac.uk/resources/downloads/

Additional proteomic data of *Clonorchis sinensis* were downloaded from HelmDB: http://gasser-research.vet.unimelb.edu.au/helmdb/.

We thank to Fatima Cvrčková, Ivan Čepička, Vojtěch Žárský, Jaroslav Kulda and Charlie Lotterman for their advices, suggestions and help.

The work was supported by Project UNCE 204004 (Charles University in Prague).

Supplementary Materials:

Supplementary Materials:

1. Organisms

Predicted proteomes – whole sets of proteins of given organisms – were obtained from the NCBI GenBank database and from the Sanger Institute. To provide sufficient length for a peptide vocabulary usage assay, only the organisms with proteome larger than 1MB (size of a briefly annotated FASTA formated file) were included in the study.

The complete list of species is available in supplementary table S1.

This study was strongly limited by the availability of proteomes of sufficient size. Some desirable proteomes are not presently available or they are too short to be included in the analysis. Also, we tried to include only those species that are unambiguously parasitic or free-living. Similar problem with the classification of organisms arises with uncertainties concerning the cellularity of organisms. We classified colonial *Volvox carteri* or *Dictyostelium discoideum* and *Polysphondylium pallidum* forming cellular slime molds as unicellular. We did not include prokaryotic organisms in this study due to their completely different status in evolution and different genomic structure as compared to eukaryotes. We also excluded fungi species because most of them have either short proteomes or are difficult to classify according to our criteria of parasitism (such as potato late blight *Phytophthora infestans* and other plant pathogens, and enthomopathogenic ectoparasites such as genus *Metarhyzium).*

2. Data filtration and standardization

Homologs and paralogs, i.e. proteins of common origin differing from each other in only limited number of amino acids are present in all eukaryotic proteomes. Only one representative of such protein family was retained and all others were excluded from analyzed proteome to avoid an artificial decrease of vocabulary size in homologs and paralogs-rich proteomes during our data sampling step, see below. Similarly, comments, annotations, and special characters occurring in sequences (coding, e.g., unknown amino acids or gaps) were filtered out during the loading procedure. Although they occur only rarely in the proteomes, they would cause a pronounced artificial inflation of alphabet size – the parameter that has a substantial effect on the size of vocabulary usage.

For data processing and all peptide frequency computations we developed our own software: "Complexity G" (*16*), which is available at figshare http://dx.doi.org/10.6084/m9.figshare.1534499. To perform particular proteome filtration, proteins were read one by one from the input text file. Our computer program randomly selected *k* peptides of length *n* from each input protein and compared these peptides with all previously read proteins. If at least one matching peptide was found then the protein was considered as a homolog or paralog of a previously read protein and was excluded from the filtrated proteome. The default length of compared peptides (*n*) was set to16 and the number of selected samples per protein (*k*) was set to 5 for most organisms. Organisms with many homologs and paralogs, such as plants, required a higher *k* – up to 20; otherwise the filtration was not strict enough. It was possible to directly verify the correct parameter settings by visual inspection of the graphical representation of vocabulary usage, as the vocabulary usage of 16-peptides (or longer k-mers as explained above) approached 1 in non-redundant (sufficiently filtrated) proteomes.

3. Data analysis

Vocabulary usage (as defined further below in eq.1.) was computed for peptides of length *n* ranging from 4 to 12 among the random samples of 1 000 000 *n*-length peptides from each proteome. Thus the final size of compared proteomes was the same. The upper bound (12 amino acids) reflects the usual length of trimmed peptides that are loaded to MHC molecules (*5*).

The V*ocabulary usage* U_n of a given organisms is defined as the ratio of the actual $U_{n,a}$ vocabulary size (the number of different peptides) to the maximal combinatorially possible vocabulary size $U_{n,max}$ for peptide length *n*.

$$U_n = \frac{U_{n,a}}{U_{n,\max}} \tag{1}$$

The theoretical number of combinatorially possible peptides of length *n* ($U_{n,max}$) was computed as follows:

$$U_{n,\max} = \min\left(1\,000\,000,\, s^n\right) \tag{2}$$

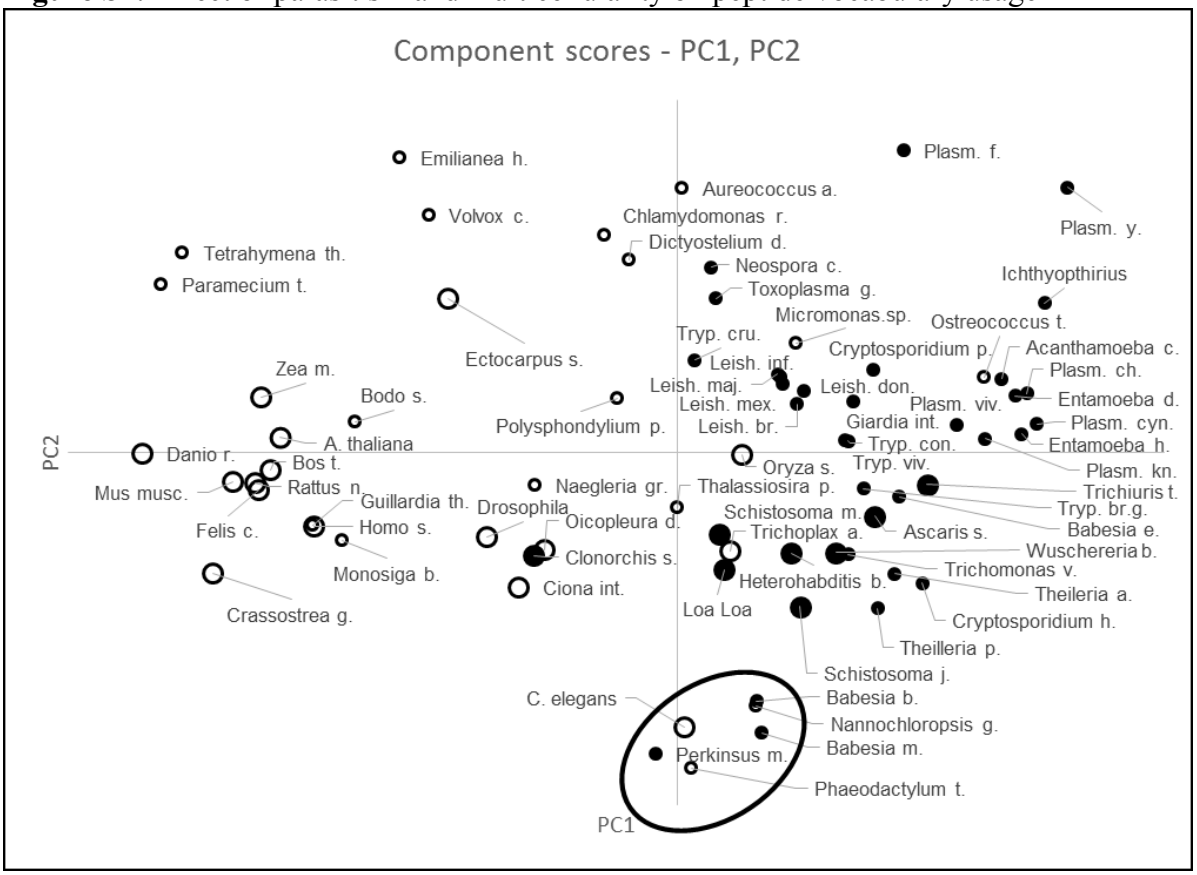
where *n* is the peptide length, and *s* is the alphabet size, i.e. 20 for amino acid alphabet.

Computation was done first for data without filtration (containing paralogs) and then for filtered data. For individual values of U_4-U_{12} see Supplementary table S3.

To reduce the risk of multiple test artefacts, we used Principal Component Analysis from a covariance matrix (unrotated) (*17*) to reduce the number of 9 dependent variables (vocabulary usage indexes for length of words from 4 to 12 amino acids) to 4 independent principle components - each explaining more than 1% of variability.

Correlations of these factors with focal binary variables were computed with ANCOVAs (type III. sum of squares). Four proteome characteristics (length of proteome before filtration, length of filtered proteome, number of proteins -unfiltered, and number of proteins -filtered) were included in the models as covariates. Because of the nested character of data (for example no parasitic autotroph exists), we could not include all focal variables and all organisms into one complex model. Instead, we first computed a basic model containing only confounding variables and then subtracted the amount of variance in vocabulary usage explained by this model from the amount of variance explained by models containing the confounding variables and one focal binary variable: namely parasitism, unicellular parasitism, multicellular parasitism, multicellularity and heterotrophy, respectively.

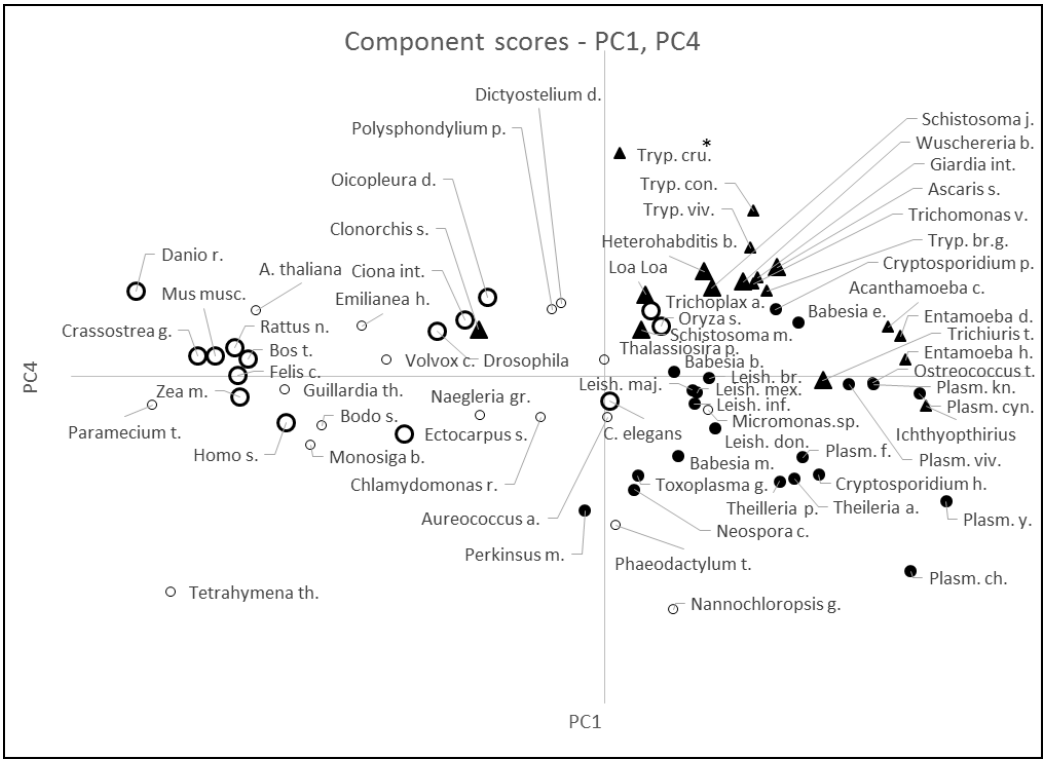
Figure S1: Effect of parasitism and multicellularity on peptide vocabulary usage



Parasitic organisms (dark circles) occupy the right part of the graph – the positive values of PC1, while free-living organisms (white circles) are located mostly in the left part of the graph. Unicellular organisms (small circles) mostly contribute to positive values of PC2 (generally poor vocabulary) while multicellular organisms

(larger circles) have negative values of PC2. The small group of organisms at the bottom of the graph (indicated by ellipse) consists exclusively of organisms with the shortest proteomes, with lengths from 1,2MB to 1,7MB. These organisms made a separate cluster because their vocabulary usage is relatively richer compared to that of species with a longer proteome. Although their actual vocabulary is poorer, it occupies a larger part of combinatorially possible words, so their vocabulary usage (as defined in eq.1) is richer.

Figure S2: Effect of parasitism and intracellularity on peptide vocabulary usage



Two-dimensional space defined by two principle components correlated with parasitism (PC1 and PC4). Dark and white circles denote the position of parasitic and free-living organisms, respectively. Triangles indicate extracellular parasites. Larger objects indicate multicellular organisms. The area of positive values of both PC1 and PC4 is occupied mostly by extracellular parasites (both multicellular and unicellular – except of intracellular parasites *Babesia bovis, B. equi.* and *Cryptosporidium parvum*. * *Trypanosoma cruzi* occupies both during its life cycle – intracellular and extracellular environment (*18*). Our results show that it has rather extracellular-like vocabulary usage.

Table S1: Effects of proteome size on peptide vocabulary usage

Length factors	PC1			PC2			PC3			PC4		
	beta	%	p-value	beta	%	p-value	beta	%	p-value	beta	%	p-value
Length of unfiltered proteome	4.59E-09	0.2	0.487	4.07E-09	0.2	0.668	-4.27E-09	2.4	0.194	1.24E-09	0.3	0.621
Length of filtered proteome	-3.49E-08	3.9	**0.002**	9.84E-09	0.5	0.525	3.61E-09	0.6	0.498	-5.34E-09	2.3	0.193
Number of proteins (unfiltered data)	-2.52E-06	0.3	0.395	-1.45E-06	0.2	0.734	2.15E-06	3.0	0.147	-4.04E-07	0.2	0.720
Number of proteins (filtered data)	3.03E-06	0.1	0.541	-1.62E-06	0.1	0.820	-2.54E-06	1.5	0.302	2.60E-06	2.6	0.170

The table summarizes the results of analyses of four simple multivariate ANCOVA models containing all four independent variables (column 1) and one dependent variable i.e. the component PC1, PC2, PC3 or PC4. The beta value computed by ANCOVA indicates size and direction of the effects of four parameters characterizing the size of the proteome on particular principal components. P-value is a two-sided statistical significance. Significant p-values are printed in bold.

Table S2: Table of species

Species name	Supergroup	Phylum	Parasite	Host (transmitters in parentheses)	Multicellularity	Autotrophism	Intracellular parasitism	Abbreviation	Source
Acanthamoeba castellanii	Amoebozoa	Acanthamoebidae	1	man	0	0	0	Acanthamoeba c.	GeneBank
Ascaris suum	Opisthokonts	Nematoda	1	pig	1	0	0	Ascaris s.	GeneBank
Babesia equi	Sar	Apicomplexa	1	horse (ixodida)	0	0	1	Babesia e.	GeneBank
Babesia bovis	Sar	Apicomplexa	1	cattle (ixodida)	0	0	1	Babesia b.	GeneBank
Babesia microti	Sar	Apicomplexa	1	man (Ixodida)	0	0	1	Babesia m.	GeneBank
Clonorchis sinensis	Opisthokonts	Platyhelminthes	1	man, felids (gastropoda)	1	0	0	Clonorchis s.	GeneBank
Cryptosporidium hominis	Sar	Apicomplexa	1	man, mammals	0	0	1	Cryptosporidium h.	GeneBank
Cryptosporidium parvum	Sar	Apicomplexa	1	man, mammals	0	0	1	Cryptosporidium p.	GeneBank
Entamoeba histolytica	Amoebozoa	Archamoeba	1	man	0	0	0	Entamoeba h.	GeneBank
Entamoeba dispar	Amoebozoa	Archamoeba	0	man	0	0	0	Entamoeba d.	GeneBank
Heterohabditis bacteriophora	Opisthokonts	Nematoda	1	insect	1	0	0	Heterohabditis b.	Sanger Institute
Giardia intestinalis	Excavata	Fornicata	1	man, mammals	0	0	0	Giardia int.	GeneBank
Ichthyophthirius multifiliis	Sar	Ciliates	1	fish	0	0	0	Ichthyophthirius	GeneBank
Leishmania brasiliensis	Excavata	Euglenozoa*	1	man, mammals (diptera)	0	0	1	Leish. br.	GeneBank
Leishmania donovanii	Excavata	Euglenozoa*	1	man, mammals (diptera)	0	0	1	Leish. don.	GeneBank
Leishmania infantum	Excavata	Euglenozoa*	1	man, mammals (diptera)	0	0	1	Leish. inf.	GeneBank
Leishmania major	Excavata	Euglenozoa*	1	man, mammals (diptera)	0	0	1	Leish. maj.	GeneBank
Leishmania mexicana	Excavata	Euglenozoa*	1	man, mammals (diptera)	0	0	1	Leish. mex.	GeneBank
Loa Loa	Opisthokonts	Nematoda	1	man, primates (diptera)	1	0	0	Loa Loa	GeneBank
Neospora caninum	Sar	Apicomplexa	1	canines (endothermic vertebrates)	0	0	1	Neospora c.	GeneBank
Perkinsus marinus	Sar	Dinoflagellata	1	oyster	0	0	1	Perkinsus m.	GeneBank
Plasmodium cynomolgi	Sar	Apicomplexa	1	monkeys (diptera)	0	0	1	Plasm. cyn.	GeneBank
plasmodium chabaudi	Sar	Apicomplexa	1	rodents (diptera)	0	0	1	Plasm. ch.	GeneBank
plasmodium falciparum	Sar	Apicomplexa	1	man (diptera)	0	0	1	Plasm. f.	GeneBank
plasmodium knowlesi	Sar	Apicomplexa	1	monkeys (diptera)	0	0	1	Plasm. kn.	GeneBank
Plasmodium vivax	Sar	Apicomplexa	1	man (diptera)	0	0	1	Plasm. viv.	GeneBank
plasmodium yoelii	Sar	Apicomplexa	1	rodents (diptera)	0	0	1	Plasm. y.	GeneBank
Schistosoma japonicum	Opisthokonts	Platyhelminthes	1	man, mammals (gastropoda)	1	0	0	Schistosoma j.	GeneBank
Schistosoma mansoni	Opisthokonts	Platyhelminthes	1	man, primates (gastropoda)	1	0	0	Schistosoma m.	GeneBank
Theileria annulata	Sar	Apicomplexa	1	livestock (Ixodida)	0	0	1	Theileria a.	GeneBank
Theilleria parva	Sar	Apicomplexa	1	livestock (Ixodida)	0	0	1	Theilleria p.	GeneBank
Toxoplasma gondii	Sar	Apicomplexa	1	felids (endothermic vertebrates)	0	0	1	Toxoplasma g.	GeneBank
Trichomonas vaginalis	Excavata	Parabasalia	1	man	0	0	0	Trichomonas v.	GeneBank
Trichiuris trichiura	Opisthokonts	Nematoda	1	man, great apes (gastropoda)	1	0	0	Trichiuris t.	GeneBank
Trypanosoma brucei gambiense	Excavata	Euglenozoa*	1	man., mammals (diptera - tsetse fly)	0	0	0	Tryp. br.g.	GeneBank
Trypanosoma congolense	Excavata	Euglenozoa*	1	livestock (diptera - tsetse fly)	0	0	0	Tryp. con.	GeneBank
Trypanosoma cruzi	Excavata	Euglenozoa*	1	man, mammals (Hemiptera -triatominae)	0	0	0/1	Tryp. cru.	GeneBank
Trypanosoma vivax	Excavata	Euglenozoa*	1	livestock (diptera - tsetse fly)	0	0	0	Tryp. viv.	GeneBank
Wuchereria bancrofti	Opisthokonts	Nematoda	1	man, great apes (diptera)	1	0	0	Wuschereria b.	GeneBank
Arabidopsis thaliana	Archaeplastida	Viridiplantae	0	NA	1	1	NA	A.thaliana	GeneBank
Chlamydomonas reinhardtii	Archaeplastida	Viridiplantae	0	NA	1	1	NA	Chlamydomonas r.	GeneBank
Micromonas sp.	Archaeplastida	Viridiplantae	0	NA	0	1	NA	Micromonas.sp.	GeneBank
Oriza sativa	Archaeplastida	Viridiplantae	0	NA	1	1	NA	Oriza s.	GeneBank
Zea mays	Archaeplastida	Viridiplantae	0	NA	1	1	NA	Zea m.	GeneBank
Bos taurus	Opisthokonts	Vertebrata	0	NA	1	0	NA	Bos t.	GeneBank
Caenorhabditis elegans	Opisthokonts	Nematoda	0	NA	1	0	NA	C. elegans	GeneBank
Ciona intestinalis	Opisthokonts	Tunicata	0	NA	1	0	NA	Ciona int.	GeneBank
Danio rerio	Opisthokonts	Vertebrata	0	NA	1	0	NA	Danio r.	GeneBank
Drosophila melanogaster	Opisthokonts	Arthropoda	0	NA	1	0	NA	Drosophila	GeneBank
Felis catus	Opisthokonts	Vertebrata	0	NA	1	0	NA	Felis c.	GeneBank
Homo sapiens	Opisthokonts	Vertebrata	0	NA	1	0	NA	Homo s.	GeneBank
Mus musculus	Opisthokonts	Vertebrata	0	NA	1	0	NA	Mus musc.	GeneBank
Oicopleura dioica	Opisthokonts	Tunicata	0	NA	1	0	NA	Oicopleura d.	GeneBank
Rattus norvegicus	Opisthokonts	Vertebrata	0	NA	1	0	NA	Ratus n.	GeneBank
Crassostrea gigas	Opisthokonts	Mollusca	0	NA	1	0	NA	Crassostrea g.	GeneBank
Aureococcus anophagefferens	Sar	Stramenopiles	0	NA	0	1	NA	Aureococcus a.	GeneBank
Bodo saltans	Excavata	Euglenozoa*	0	NA	0	0	NA	Bodo s.	Sanger Institute
*Dictyostelium discoideum***	Amoebozoa	Discosea	0	NA	0	0	NA	Dictyostelium d.	GeneBank
Ectocarpus siliculosus	Sar	Stramenopiles	0	NA	1	1	NA	Ectocarpus s.	GeneBank
Emiliania Huxleyii	Chromalveolata	Haptophyta	0	NA	0	1	NA	Emiliania h.	GeneBank
Guillardia theta	Chromalveolata	Cryptophyta	0	NA	0	1	NA	Guillardia th.	GeneBank
Monosiga brevicollis	Opisthokonts	Choanoflagellate	0	NA	0	0	NA	Monosiga b.	GeneBank
Naegleria gruberi	Excavata	Heterolobosea	0	NA	0	0	NA	Naegleria gr.	GeneBank
Nannochloropsis gaditana	Sar	Stramenopiles	0	NA	0	1	NA	Nannochloropsis g.	GeneBank
Ostreococcus tauri	Archaeplastida	Viridiplantae	0	NA	0	1	NA	Ostreococcus t.	GeneBank
Paramecium tetraurelia	Sar	Ciliates	0	NA	0	0	NA	Paramecium t.	GeneBank
Phaeodactylum tricornutum	Sar	Stramenopiles	0	NA	0	1	NA	Phaeodactylum t.	GeneBank
*Polysphondylium pallidum***	Amoebozoa	Discosea	0	NA	0	0	NA	Polysphondylium p.	GeneBank
Tetrahymena thermophila	Sar	Ciliates	0	NA	0	1	NA	Tetrahymena th.	GeneBank
Thalassiosira pseudonana	Sar	Stramenopiles	0	NA	0	1	NA	Thalassiosira p.	GeneBank
Trichoplax adhaerens	Opisthokonts	Placozoa	0	NA	1	0	NA	Trichoplax a.	GeneBank
*Volvox carteri****	Archaeplastida	Viridiplantae	0	NA	0	1	NA	Volvox c.	GeneBank

* class Kinetoplastida
*organisms forming cellular slime molds
**organism forming a colony
NA = not applicable

Complete list of 72 eukaryotic species. Organisms were classified according to the last revision of eukaryotes made by Adl and colleagues (*19) E. dispar* is not a parasite but a commensal of the human gut which is almost morphologically identical to the pathogenic *E. histolytica* (*20, 21*). Although it is not invasive, its tendency to hide from immunity systems is the same as with pathogenic parasites, so its position close to *E. histolytica* in the component score space is reasonable. Although *Tetrahymena thermophila* is able to switch to a parasitic mode of survival, we included it into the study as a free-living unicellular organism, because its parasitic tendency seems to be quite rare (*22*).

Table S3: Individual values of U_4-U_{12}

Organism	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}	U_{11}	U_{12}
Acanthamoeba castellanii	0.973806	0.434357	0.853636	0.956815	0.977540	0.984417	0.987822	0.989884	0.991281
Ascaris suum	0.987369	0.519373	0.891412	0.971668	0.981647	0.984951	0.986983	0.988474	0.989593
Babesia equi	0.946200	0.507762	0.891728	0.952634	0.964733	0.970937	0.975319	0.978638	0.981232
Babesia bovis	0.944619	0.619843	0.945723	0.981758	0.985785	0.987264	0.988344	0.989226	0.989985
Babesia microti	0.927700	0.629147	0.953603	0.990733	0.994129	0.994911	0.995361	0.995691	0.995948
Clonorchis sinensis	0.995350	0.649224	0.834400	0.961671	0.976926	0.981319	0.984107	0.986221	0.987899
Cryptosporidium hominis	0.925331	0.527885	0.917605	0.98362	0.991255	0.993340	0.994539	0.995417	0.996078
Cryptosporidium parvum	0.970050	0.476346	0.826767	0.946834	0.963430	0.968895	0.972600	0.975520	0.977879
Entamoeba histolytica	0.952556	0.445649	0.880170	0.967398	0.979260	0.982876	0.985059	0.986675	0.987933
Entamoeba dispar	0.951819	0.436216	0.866121	0.956105	0.969817	0.974698	0.977923	0.980423	0.982447
Heterohabditis bacteriophora	0.990694	0.559313	0.89043	0.971525	0.981531	0.985138	0.987631	0.989578	0.991125
Giardia intestinalis	0.967994	0.493232	0.843134	0.938115	0.955265	0.963787	0.970045	0.97493	0.978769
Ichthyopthirius multifiliis	0.943363	0.398373	0.825634	0.947328	0.968270	0.975310	0.979703	0.982937	0.985458
Leishmania brasiliensis	0.985188	0.512256	0.814862	0.963223	0.984944	0.988337	0.989312	0.989840	0.990225
Leishmania donovanii	0.984119	0.506180	0.805539	0.964342	0.991388	0.995965	0.997091	0.997559	0.997836
Leishmania infantum	0.984719	0.510003	0.797826	0.957596	0.984996	0.989701	0.990924	0.991476	0.991832
Leishmania major	0.9846	0.510053	0.797682	0.955633	0.982285	0.986913	0.988139	0.9887	0.98907
Leishmania mexicana	0.984850	0.511118	0.802302	0.958411	0.984124	0.988469	0.989635	0.990193	0.990569
Loa Loa	0.992394	0.587527	0.881511	0.973178	0.984037	0.9874	0.98942	0.990876	0.991953
Neospora caninum	0.978656	0.501137	0.729557	0.939309	0.980194	0.987536	0.989674	0.990774	0.991536
Perkinsus marinus	0.900406	0.674434	0.942842	0.976210	0.981494	0.983832	0.985479	0.986809	0.987897
Plasmodium cynomolgi	0.951256	0.437468	0.875154	0.970352	0.983654	0.987628	0.990009	0.991746	0.993115
plasmodium chabaudi	0.902575	0.435387	0.849108	0.966001	0.983362	0.986949	0.988673	0.989879	0.990814
plasmodium falciparum	0.937925	0.401231	0.730795	0.908140	0.948555	0.960250	0.966614	0.971222	0.974878
plasmodium knowlesi	0.956963	0.459475	0.870928	0.971464	0.984145	0.987335	0.989122	0.990424	0.991449
Plasmodium vivax	0.956794	0.465167	0.860309	0.965303	0.980421	0.984557	0.986850	0.988463	0.989728
plasmodium yoelii	0.920856	0.356982	0.774462	0.930492	0.962015	0.970731	0.975574	0.979101	0.981893
Schistosoma japonicum	0.958550	0.574744	0.920571	0.966639	0.973145	0.975792	0.977652	0.979149	0.980405
Schistosoma mansoni	0.989356	0.578345	0.861152	0.971566	0.985174	0.988404	0.990002	0.991061	0.991846
Theileria annulata	0.914756	0.535359	0.910959	0.972870	0.981868	0.985463	0.987877	0.989706	0.991109
Theilleria parva	0.915213	0.551635	0.921862	0.978089	0.985589	0.988276	0.990001	0.991288	0.992294
Toxoplasma gondii	0.980819	0.508920	0.744970	0.944367	0.981606	0.988874	0.991318	0.992619	0.993492
Trichomonas vaginalis	0.932642	0.495437	0.887349	0.955243	0.967596	0.973414	0.977489	0.980603	0.983059
Trichiuris trichiura	0.989813	0.539935	0.899864	0.979011	0.987927	0.990645	0.992313	0.993526	0.994461
Trypanosoma brucei gambiense	0.986950	0.514694	0.873349	0.973171	0.983614	0.985579	0.986540	0.987213	0.987737
Trypanosoma congolense	0.986719	0.505226	0.860631	0.951968	0.963231	0.966949	0.969575	0.971719	0.973514
Trypanosoma cruzi	0.988406	0.533456	0.804657	0.917151	0.934554	0.941525	0.946794	0.951310	0.955289
Trypanosoma vivax	0.985388	0.506381	0.855535	0.955596	0.967867	0.971764	0.974530	0.976830	0.978789
Wuchereria bancrofti	0.988738	0.543807	0.897308	0.977428	0.986393	0.988896	0.990313	0.991318	0.992074
Arabidopsis thaliana	0.996869	0.701460	0.741265	0.909451	0.940184	0.952887	0.961518	0.968047	0.973165
Chlamydomonas reinhardtii	0.989388	0.527704	0.707165	0.903725	0.957812	0.974488	0.981210	0.984653	0.986758
Oryza sativa	0.924681	0.550999	0.851022	0.913861	0.934104	0.946021	0.954531	0.961112	0.966339
Zea mays	0.997756	0.695887	0.710918	0.908197	0.951051	0.965061	0.972823	0.978229	0.982293
Micromonas sp.	0.982263	0.494215	0.787247	0.949719	0.980251	0.987387	0.990266	0.991942	0.993082
Bos taurus	0.997881	0.714787	0.746096	0.924846	0.954520	0.964113	0.970061	0.974505	0.977994
Caenorhabditis elegans	0.936688	0.653808	0.94224	0.977412	0.983203	0.985706	0.987285	0.988428	0.989306
Ciona intestinalis	0.9958	0.664155	0.846231	0.964385	0.978336	0.982135	0.984375	0.986003	0.987275
Danio rerio	0.998338	0.753955	0.723564	0.901914	0.932083	0.943392	0.951029	0.956973	0.961771
Drosophila melanogaster	0.995438	0.660165	0.818451	0.951457	0.971965	0.978207	0.981563	0.983744	0.985280
Felis catus	0.998194	0.724847	0.749784	0.930351	0.959058	0.968039	0.973507	0.977571	0.980727
Homo sapiens	0.997881	0.716768	0.769699	0.947954	0.975152	0.982519	0.986507	0.989250	0.991253
Mus musculus	0.998238	0.731161	0.744108	0.923875	0.952840	0.962337	0.968314	0.972828	0.976376
Oicopleura dioica	0.994444	0.643802	0.838361	0.956875	0.97152	0.975929	0.978687	0.980783	0.982474
Rattus norvegicus	0.998019	0.723430	0.749611	0.924628	0.952833	0.962311	0.968356	0.972941	0.976561
Crassostrea gigas	0.998431	0.765783	0.77818	0.940363	0.962062	0.968844	0.973145	0.976417	0.979039
Aureococcus anophagefferens	0.983169	0.487158	0.702295	0.902164	0.957012	0.971954	0.977470	0.980395	0.982363
Bodo saltans	0.997025	0.670798	0.732913	0.933352	0.968907	0.977302	0.981094	0.983497	0.985298
Dictyostelium discoideum	0.976244	0.527361	0.737576	0.896705	0.932189	0.947259	0.957115	0.964284	0.969699
Ectocarpus siliculosus	0.992931	0.601257	0.700828	0.909851	0.958369	0.972955	0.979752	0.983798	0.986480
Emiliania Huxleyii	0.993244	0.574496	0.644633	0.868615	0.926473	0.941776	0.948507	0.952901	0.956271
Guillardia theta	0.998056	0.716691	0.771772	0.947938	0.971468	0.976810	0.979649	0.981698	0.983300
Monosiga brevicollis	0.997588	0.711385	0.777209	0.959982	0.985392	0.989714	0.991232	0.992120	0.992746
Naegleria gruberi	0.988656	0.628012	0.796642	0.956327	0.980290	0.986698	0.989864	0.991817	0.993174
Nannochloropsis gaditana	0.883300	0.626319	0.932895	0.983253	0.990945	0.993029	0.994062	0.994710	0.995181
Ostreococcus tauri	0.970094	0.440274	0.841656	0.967709	0.986268	0.989695	0.990987	0.9918	0.992408
Paramecium tetraurelia	0.991525	0.696312	0.642256	0.887872	0.932588	0.945477	0.953593	0.959924	0.965076
Phaeodactylum tricornutum	0.905300	0.666350	0.952825	0.986654	0.990864	0.992405	0.993430	0.994209	0.994828
Polysphondylium pallidum	0.984656	0.572965	0.791259	0.925321	0.952750	0.964548	0.971957	0.977099	0.980761
Tetrahymena thermophila	0.987900	0.680145	0.610725	0.906169	0.964513	0.975429	0.979964	0.982995	0.985328
Thalassiosira pseudonana	0.991231	0.585206	0.838189	0.968218	0.985719	0.989657	0.991534	0.992789	0.993706
Trichoplax adhaerens	0.989938	0.579881	0.873449	0.971202	0.983367	0.987319	0.989766	0.991499	0.992774
Volvox carteri	0.993263	0.582311	0.672604	0.876925	0.933317	0.953487	0.963611	0.969634	0.973527

The table demonstrates certain vocabulary usages for peptides of length 4 to 12 of particular organisms. As PCA showed, the vocabularies of individual organisms differ mostly in values of U_4-U_6 and then the index converge to one.

Table S4: Complete results of ANCOVA

	PC1			PC2			PC3			PC4		
	beta	%	p-value	beta	%	p-value	beta	%	p-value	beta	%	p-value
Parasitism– Non-parasitism	6.61E-02	6.5	**0.000**	3.32E-02	2.8	0.137	-1.58E-02	5.7	**0.041**	7.71E-03	2.3	0.196
length of unfiltered proteome	6.06E-09	0.3	0.305	3.57E-09	0.2	0.704	-5.07E-09	3.2	0.121	1.11E-09	0.3	0.660
length of filtered proteome	-3.12E-08	3.0	**0.002**	1.32E-08	1.0	0.386	3.15E-09	0.5	0.550	-4.65E-09	1.7	0.256
number of proteins (unfiltered data)	-3.36E-06	0.5	0.205	-1.44E-06	0.1	0.733	2.48E-06	3.9	0.091	-4.03E-07	0.2	0.721
number of proteins (filtered data)	4.51E-06	0.3	0.306	-1.19E-06	0.0	0.865	-2.96E-06	2.0	0.223	2.72E-06	2.8	0.150
R^2	0.813			0.171			0.132			0.125		
Unicellular parasitism.– Unicell. non-parasitism	5.67E-02	5.7	**0.007**	3.45E-02	2.4	0.220	-3.75E-03	0.3	0.680	9.50E-03	2.7	0.260
length of unfiltered proteome	1.44E-08	1.3	0.173	1.02E-08	0.8	0.490	-1.02E-08	7.9	**0.038**	2.84E-09	0.9	0.520
length of filtered proteome	-4.24E-08	4.8	**0.011**	4.99E-09	0.1	0.825	3.92E-09	0.5	0.595	-8.46E-09	3.2	0.216
number of proteins (unfiltered data)	-6.65E-06	1.2	0.195	-1.96E-06	0.1	0.782	5.79E-06	10.8	**0.016**	-2.91E-07	0.0	0.892
number of proteins (filtered data)	1.04E-05	1.2	0.188	3.49E-06	0.2	0.749	-3.76E-06	1.9	0.295	3.58E-06	2.5	0.278
R^2	0.717			0.355			0.301			0.165		
Multicell. par. – Multicell. non-parasitism	6.27E-02	5.3	**0.001**	-1.90E-02	1.6	0.545	-2.96E-02	16.4	0.057	8.09E-03	6.2	0.190
length of unfiltered proteome	5.05E-09	0.3	0.348	2.31E-09	0.2	0.822	-6.25E-09	6.8	0.208	-3.03E-09	8.1	0.137
length of filtered proteome	-3.45E-08	5.7	**0.001**	-2.49E-09	0.1	0.881	7.46E-09	3.7	0.349	3.19E-09	3.4	0.326
number of proteins (unfiltered data)	-2.81E-06	0.6	0.218	-1.43E-06	0.5	0.741	2.74E-06	7.4	0.189	1.05E-06	5.5	0.215
number of proteins (filtered data)	5.32E-06	0.7	0.170	4.82E-06	1.9	0.512	-4.77E-06	7.8	0.177	-8.90E-07	1.4	0.530
R^2	0.935			0.192			0.245			0.362		
Unicellularity – Multicellularity	4.09E-02	2.6	**0.008**	7.61E-02	15.6	**0.000**	-1.20E-03	0.0	0.877	-1.08E-02	4.6	0.064
length of unfiltered proteome	7.50E-09	0.5	0.252	7.30E-09	0.8	0.403	-4.95E-09	3.0	0.145	4.34E-10	0.0	0.862
length of filtered proteome	-4.07E-08	5.1	**0.000**	9.66E-10	0.0	0.946	4.27E-09	0.9	0.439	-3.67E-09	1.1	0.370
number of proteins (unfiltered data)	-3.60E-06	0.5	0.219	-2.72E-06	0.5	0.485	2.36E-06	3.5	0.120	-1.06E-07	0.0	0.924
number of proteins (filtered data)	6.58E-06	0.6	0.189	4.62E-06	0.5	0.488	-2.72E-06	1.6	0.293	1.62E-06	0.9	0.396
R^2	0.774			0.298			0.075			0.149		
Unicell. parasitism – Multicell. parasitism	4.81E-02	6.5	0.105	1.24E-01	17.7	**0.002**	4.39E-03	0.2	0.759	2.93E-03	0.1	0.825
length of unfiltered proteome	-3.05E-09	0.2	0.789	-1.36E-08	1.4	0.340	-4.82E-09	1.5	0.391	-3.85E-09	1.3	0.459
length of filtered proteome	-1.14E-08	0.9	0.542	5.43E-08	8.3	**0.025**	-3.45E-09	0.3	0.707	5.22E-09	0.9	0.540
number of proteins (unfiltered data)	8.86E-07	0.1	0.859	5.83E-06	1.3	0.352	2.11E-06	1.5	0.390	2.40E-06	2.5	0.295
number of proteins (filtered data)	2.47E-06	0.3	0.720	-4.23E-06	0.4	0.623	-3.17E-06	1.8	0.351	1.22E-07	0.0	0.969
R^2	0.251			0.518			0.366			0.284		
Unicell. non-parasitism– Multicell. non-parasitism	4.12E-02	3.8	**0.025**	7.09E-02	14.5	**0.022**	-7.61E-03	1.5	0.493	-1.25E-02	9.5	0.076
length of unfiltered proteome	1.10E-08	1.5	0.148	8.84E-09	1.3	0.482	-5.20E-09	4.0	0.272	8.02E-10	0.2	0.783
length of filtered proteome	-4.36E-08	8.4	**0.001**	-1.17E-08	0.8	0.578	4.90E-09	1.3	0.532	-4.52E-09	2.5	0.356
number of proteins (unfiltered data)	-5.62E-06	2.0	0.098	-4.30E-06	1.5	0.443	2.64E-06	5.2	0.212	-5.99E-07	0.6	0.645
number of proteins (filtered data)	1.04E-05	2.0	0.094	1.14E-05	3.2	0.266	-3.40E-06	2.6	0.374	2.60E-06	3.5	0.276
R^2	0.812			0.307			0.112			0.212		
Heterotrophy – Autotrophy	2.17E-02	0.7	0.189	-3.74E-02	3.4	0.103	-9.20E-03	1.8	0.255	1.65E-02	9.8	**0.006**
length of unfiltered proteome	4.95E-09	0.2	0.464	3.77E-09	0.2	0.687	-4.74E-09	2.8	0.156	7.54E-10	0.1	0.754
length of filtered proteome	-3.50E-08	3.9	**0.002**	1.09E-08	0.6	0.473	4.03E-09	0.8	0.453	-4.96E-09	2.0	0.204
number of proteins (unfiltered data)	-2.59E-06	0.3	0.392	-1.39E-06	0.1	0.740	2.27E-06	3.2	0.129	-2.13E-07	0.0	0.843
number of proteins (filtered data)	3.29E-06	0.2	0.513	-2.30E-06	0.1	0.741	-2.71E-06	1.7	0.274	2.72E-06	2.8	0.132
R^2	0.755			0.176			0.093			0.201		
Intracelullar parasitism– Extracellular parasitism	-1.21E-02	0.7	0.585	2.07E-02	0.8	0.495	1.15E-02	2.2	0.275	-2.82E-02	17.4	**0.004**
length of unfiltered proteome	1.16E-09	0.0	0.918	-7.74E-09	0.5	0.616	-3.98E-09	1.0	0.459	1.18E-09	0.1	0.802
length of filtered proteome	-8.80E-09	0.5	0.653	5.95E-08	8.9	**0.032**	-7.32E-09	1.1	0.432	4.96E-10	0.0	0.951
number of proteins (unfiltered data)	-1.61E-06	0.3	0.740	1.70E-06	0.1	0.797	2.14E-06	1.6	0.354	6.95E-08	0.0	0.973
number of proteins (filtered data)	3.38E-07	0.0	0.961	-7.66E-06	1.2	0.424	-2.42E-06	1.0	0.467	4.57E-07	0.0	0.875
R^2	0.117			0.477			0.437			0.463		

The table summarizes the results of analyses of four simple multivariate ANCOVA models containing all five independent variables (column 1) and one dependent component (PC1-4). The beta value computed by ANCOVA indicates the size and direction of effects of five parameters characterizing the size of the proteome on particular principal components. P- value is a two-sided statistical significance of corresponding beta.

Coefficient of determination R^2 indicates the total variability explained by the model. Variability proportions of individual factors were obtained from type III sum of squares.

Table S5: Comparison of 5 clades containing both-parasitic and free-living organisms for values of PC1 a PC3

Clade	Opisthokonts		Kinetoplastids		Ciliates		Nematods		Sar	
	parasites	free-living (multicell.)	parasites	free-living	parasites	free-living	parasites	free-living	parasites	free-living
PC1	0.0404	-0.1190	0.0501	-0.1324	0.1508	-0.2075	0.0193	0.0027	0.0864	0.0271
PC3	-0.0114	-0.0015	-0.0156	-0.0197	0.0126	0.0050	-0.0263	0.0209	0.0098	-0.0051

Average values of PC1 and PC3, p-value for exact binomial test are 0.031 for 5 pairs and 0.0009 for 10 compared pairs.

2. Zemkova M, Trifonov EN, Zahradnik D. (2014) One common structural feature of "words" in protein sequences and human texts. J Biomol Struct Dyn;32(7):1085-91.

M. Zemková[a]*, E.N. Trifonov[b] and D. Zahradník[c]
[a]Faculty of Science, Department of Philosophy and History of Science, Charles University in Prague, Viničná 7, Praha CZ-12844,Czech Republic;
[b]Genome Diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel;
[c]Faculty
of Forestry and Wood Sciences, Department of Forest Management, Czech University of Life Sciences Prague, Kamýcká 1176, Praha,CZ-165 21, Czech Republic
Communicated by Ramaswamy H. Sarma
(Received 11 March 2013; final version received 24 May 2013)

Abstract

Frequently discussed analogy between genetic and human texts is explored by comparison of alternation of polar and non-polar amino-acid residues in proteins and alternation of consonants and vowels in human texts. In human languages, the usage of possible combinations of consonants and vowels is influenced by pronounceability of the combinations. Similarly, oligopeptide composition of proteins is influenced by requirements of protein folding and stability. One special type of structure often present in proteins is amphipathic α-helices in which polar and non-polar amino acids alternace with the period 3.5 residues, not unlike alternation of consonants and vowels. In this study, we evaluated the contribution made by amphipathic alternations to the protein sequence texts (20–24%). Their proportion is lower than respective values for alternating words in human texts (57–89%). The proteomes (full sets of proteins for selected organisms) were transformed into ranked sequences of n-grams (words of length n), including periodical amphipathic structures. Similarly, human texts were transformed into sequences of alternating consonants and vowels. Analysis of the vocabularies shows that in both types of texts (human languages and proteins) the alternating words are dominant or highly preferred, thus, strengthening the analogy between these two types of texts. The contribution of amphipathic words in the upper parts of the ranked lists for 10 analyzed proteomes varies between 58 and 74%. In human texts respective values range between 90 and 100%.

Keywords: amphipathic α-helices; phonetics; consonants; vowels; pronounceability

1. Introduction

1.1. Linguistic analogy and *n*-gram based methods

The most common approach to genome comparison and investigation of their evolutionary and structural similarity is the alignment of one or more pairs of homologous genes. Then the result depends on choice of compared segments and on the quality of the alignment (Bolshoy,2003). The linguistic approaches used for investigation of genetic texts are an alternative to the alignment techniques. The linguistic techniques treat the sequences as combinations of "words" of length n and analyze their frequencies and their distribution in the texts. These "words" are usually called n-grams (Shannon, 1948). The n-gram based sequence decomposition is the ground method which provides the

list of oligomers (substrings of nucleotides or peptides), i.e. a vocabulary of words of selected chosen length n. Such an approach cannot replace standard alignment comparison but it can serve as an important supplementary information.

Linguistic tools have been efficiently used since the '80s. For example, Brendel and colleagues used a method of "so-called" contrast words to show that every organism exhibits a distinct vocabulary and also that linguistic properties vary within genome itself (Beckmann, Brendel, & Trifonov, 1986; Brendel, Beckmann, & Trifonov, 1986). Pietrokovski and Trifonov (1992) identified by this technique imported sequences in mitochondrial yeast genome. The term "signature" has been used in this work in the context of taxonomy and later became popular often specified as "genomic" or "proteomic signature". (In human texts this term would probably correspond to the "author's style.") A detailed inventory of linguistic tools and its usage in genomics are given in review (Bolshoy, 2003). Recently linguistic tools were sed by Ganapathiraju and colleagues for comparative genome n-gram analysis. The authors showed that various organisms exhibit typical compositions of their peptide vocabularies and some particular amino acid *n*-grams can be viewed as "proteome signatures". (Ganapathiraju et al., 2002; Osmanbeyoglu & Ganapathiraju, 2011).

The *n*-gram based analysis is normally used in natural language processing, NLP (such as speech processing, information retrieval, machine translation, and many other applications). Linguistic metaphors, such as "symbols", "words," and "text" applied to genetic sequences have, of course, their limits. Human and genetic texts are of fundamentally different nature. Besides, genetic sequences are multicode texts, carrying many superimposed

messages (Trifonov, 1989). One letter can belong to different messages simultaneously. Human texts, on the contrary, are read by only one way, letter by letter, and a wider field of interpretations appears only on semantic level (Popov, Segal, & Trifonov, 1996).

In both cases, in genetic and human texts, we have a sequence of characters of length *N*, length of words n, and size of the alphabet *S*. Natural texts, contrary to random ones, never use all possible combinations of the characters. There are, for example, some physical constraints selecting or preventing specific combinations.(Eco, 1976; Shannon, 1948). Using analogy between genetic and human texts, it would be rewarding to find something, indeed, comparable. One such exercise could be the comparison of well-known alternation of consonants and vowels in human texts and similar alternation of, say, polar and non-polar amino acids in amphipathic α-helices which have a substantial share in protein structures (Epand, 1993). In human languages, the usage of possible combinations of consonants and vowels is limited by the pronounceability of combinations of phonemes (physical constraint). Similarly, sequences of proteins are dictated by chemical-physical properties, such as polar or non-polar character of the amino-acid residues, and their balanced proportions. Only specific combinations of the residues do provide a correct folding of the protein in its functional 3-D structure. The amphipathic α-helices where the polar and non-polar residues alternate represent one important structural class in proteins (Mant, Zhou, & Hodges, 1993).

The proteome (the set of all proteins of some organism, the "text" in the linguistic analogy) contains besides the structured proteins also many fully or partially unfolded proteins (Dunker, Obradovic, Romeo, Garner, & Brown, 2000; Fink, 2005). In other words, there is certain amount of protein sequences which do not contain the amphipathic segments, while human languages have the alternating consonants and vowels in any stretch of the text longer than few words. The amphipathic α-helices have polar residues on one side of the helix and non-polar ones – on the other side, with the period of alternation ≈3.5 residues (see the Methodology). As the results show, distribution of potentially amphipathic words varies somewhat between organisms, so that amphipathicity could be considered as one of "proteomic signatures". Similarly, in the case of human languages, we can distinguish groups of languages according to their consonant/vowel usage.

Perceiving protein sequences as texts, the peptides which fulfill the conditions necessary for forming amphipathic α-helices will appear as words analogous to frequent vowel/consonant alternating words of human texts.

1.2. Alternation of polar and non-polar amino acids in proteins

The structural condition for forming the amphipathic α-helix is the alternation between polar (hydrophilic) and non-polar (hydrophobic) residues with a period ≈3.5 residues (i.e. distance between residues of the same nature should be 3 or 4). In this case, one side of the helix is polar, while the opposite side, non-polar. Typical size of the α-helices (the "word" size) is 10–15 residues (e.g.Berezovsky, Grosberg, & Trifonov, 2000). The list of polar and non-polar residues (Hausman & Cooper, 2004) is presented in the Table 1. For convenience, here and further on the non-polar residues are shown in italics. One good example of amphipathic helices is given by the dominant hidden motif found in prokaryotic protein sequences corresponding to the word EKFRKFSKIL (Rapoport & Trifonov, 2011).

1.3. Alternation of vowels and consonants in human languages

Pronounceability of combinations of consonants and vowels is an issue of phonetics and phonology (Clark & Yallop, 1995; Hall, 2005). It is important to note that our simplified model takes consonants and vowels as letters, not as sounds as it is common in the field of phonetics, for the purpose of demonstrating the analogy between pronounceability in human languages and physicalchemicalconstraints in proteins. Any universal model of phonetic transcription of the languages is problematic,and we do not deal with these aspects, by using only written texts. The "phonetic" texts, naturally, would have the same property of alternation. The ratio of consonants and vowels (letters) is different in the written texts in various languages. An extreme case of language where vowels are dominant is the group of Hawaiian languages. Indo-European languages have slightly dominant proportion of consonants. Within this group, one can distinguish also separate branches like German, Slavic or Roman, and other languages. In addition, there are languages using mostly consonants in written form and vowels are deduced from the context of consonants, such as Arabic and Hebrew. The North-African Berberic languages seem to be phonetic extremes of consonantal languages (Lehmann, 1992). In this work, we have chosen three Indo-European languages representing three different branches (Anglo-Saxons, Roman, and Slavic respectively): English, French, and Czech.

Table 1. List of polar and non-polar amino acids.

Polar amino acids	D, E, H, K, N, Q, R, S, T, Y
Non-polar amino acids	A, C, F, G, I, L, M, P, V, W

2. Methods and materials

2.1. n-gram based text decomposition

For the analysis of "words" of length n, the classical Shannon's method of n-grams was used (Shannon, 1948; Volkovich,Kirzhner, Bolshoy, Nevo, & Korol, 2005). The sequences of symbols in human languages were converted to the sequences of consonants and vowels (C and V), while the protein sequences were converted to sequences of non-polar and polar amino acid residues (N and P).

For genetic texts (proteomes), frequencies of occurrence of different n-grams of length n = 12 (the most common length of amphipathic α-helices) were calculated. The same was done with human languages for length of 4, 5, and 6, which are the typical human word sizes.

For all n-gram operations we used our software "Complexity_G" (and its modification for human texts
"Complexity_H") available at http://web.natur.cuni.cz/ filosof/index.php/en/staff/362-zemkova.html.

2.2. Sequences analyzed

The proteome sequences were downloaded from the NCBI reference database http://www.ncbi.nlm.nih.gov/ protein. Totally 32 different organisms have been taken for the analysis. Ten of them representing groups from all kingdoms are listed in the Table 2. The species are chosen to represent widest possible spectrum of linguistic properties. Viruses have not been considered as they have very small size of proteomes. To avoid possible biases, the proteomes were filtered to remove similar or almost identical proteins. (Final total lengths of proteomes in Table 2 refer to the filtered proteomes.)

Table 2. List of species.

Organism	Group	Size of the proteome (amino acids)
Bacillus subtilis	Prokaryote	1 213 559
Plasmodium yoelii	Chromista	3 184 179
Trypanosoma brucei	Excavata	4 327 922
Monosiga brevicolis	Opisthokonts	5 167 269
Dictyostelium discoideum	Amoebozoa	6 040 782
Saccharomyces cerevisiae	Opisthokonts	2 693 904
Micromonas sp.	Archaeplastida	4 887 405
Arabidopsis thaliana	Archaeplastida	9 705 145
Bos taurus	Opisthokonts	9 874 771
Homo sapiens	Opisthokonts	9 516 609

2.3. Human texts

To show linguistic analogy, we analyzed two kinds of texts: One novel (the abridged version of the story by Jean Giono: “L’homme qui plantait des arbres” (“Man who planted the trees”), French in the original, its translation in English, and its translation in Czech) and one juristic text (Universal Declaration of Human Rights by UN), in these three languages. As analyzed, these two texts do not display different properties in the sense of alternation of consonant and vowels.

For the sake of comparisons, the human texts usedfor the analysis are treated ignoring blanks, making them similar to continuous protein sequences. That, naturally, creates the “border words”, overlaps between neighboringwords proper. Since the speech is continuous, largely without separation between the words, the pronounceability would require more frequent alternation of vowels and consonants also in the “border words”. The texts in French and English were downloaded from http://www.perso.ch/arboretum/pla.htm, http://www.un.org/en/documents/udhr/.
Czech translation: http://www.stromy.unas.cz/muz_ktery_sazel_stromy.htm, http://www.osn.cz/dokumenty-osn/ soubory/vseobecna-deklarace-lidskych-prav.pdf.

2.4. Detection of amphipathic n-grams

The words (*n*-grams) fulfilling the condition for forming the structure of amphipathic α-helices were derived from the list of n-grams according to the following algorithm, essentially demanding the discrete strings of symbols to follow the 3.5 letter periodicity:

The (P, N) sequence belongs to amphipathic helices if P, PP, or PPP alternate with N, NN, or NNN, in such a way that there is such P (call its coordinate 0), that at position 3 or 4 or both another P is found, at position 7 yet another one, and at position 10 or 11 or both – one more. The same can be demanded for N.

The above example of dominant hidden amphipathic motif in prokaryotes (EKFRKFSKIL), of length 10 in this case, would read:

P P N P P N P P N N
0 1 2 3 4 5 6 7 8 9

2.5. Distribution of amphipathic n-grams

Let's have a sequence of amino acids divided into sequence of (overlapping) n-grams. The n-grams would correspond to every possible succession of polar and non-polar amino acid residues (e.g. PNNN, PNNPPP, …). Some of these n-grams are amphipathic (according to the definition above). A complete list of all possible different amphipathic and non-amphipathic n-grams of length 12 contains 4096 words (212 combinations for 2-letter alphabet). Calculation according to the algorithm described above shows that 896 of them are amphipathic, 12 g and 3200 non-amphipathic ones. Individual n-grams occur with different frequencies. For the purpose of the analysis below, we arranged them according to the frequency in descending (ranking) order. An output of the computation is the list of 12-grams. The contribution made by the amphipathic alternations to the protein sequence texts for 10 analyzed species ranges between 20 and 24%. Since the amphipathicity is one of the dominant features of the protein sequences, the amphipathic n-grams would be expected to occur more often in the upper part of the sorted list. To verify this hypothesis, first, one has to demonstrate that the distribution of amphipathic words is not uniform. To reject the hypothesis of uniformity we used Kolmogorov–Smirnov test which is based on evaluation of distribution of amphipathic n-grams among all n-grams in the sorted list of n-grams. This test is usually used for continuous data, or for a large set of data, as in our case. Let's denote X rank of amphipathic n-grams randomly selected from the sorted list. If the usage of amphipathic n-grams is random, then X would have uniform distribution along the list. Null hypothesis (of the uniform distribution) is rejected if the Kolmogorov–Smirnov test statistic D [1] exceeds the critical value $D_n(\alpha)$ [2].

$$D = \sup |F(x) - G(x)| \qquad (1)$$

where F is empirical distribution function of X and G is the distribution function of uniform distribution with correspondingparameters (Sheskin, 2004). Critical value can be approximated for large n at the level of significance

$$D_n(\alpha) = \sqrt{\frac{1}{2n}} \ln 1/\alpha \qquad (2)$$

The graph of distribution function visualizes then the identity of distribution of amphipathic n-grams with or deviation from the uniform distribution function. The test applied to the rankings for all analyzed species holds well (not shown). For example, for the proteome of Plasmodium yoelii, total about 4 million 12 g, the test statistic D is .130 whereas critical value $D_n(.05)$ is .021, thus rejecting the hypothesis of uniformity.

Second, we have to show that amphipathic words are dominant, that is, residing preferentially in the upper half of the list. Again, a simple calculation demonstrates thatthis is the case for all species studied. Indeed, for different species from 58 to 74% of all amphipathic words are residing in upper halfs of respective ranked lists. For illustration purposes we also divided the list of ranked 12-grams in 10 consecutive parts (bins), 400 ngrams each (the remaining infrequent 96 words are discarded).
For each bin the number of amphipathic words of the vocabulary is registered in corresponding histogram – see Figure 1. Resulting curves of the occurrences of the amphipathic words are expected to have a descending character.

3. Results

3.1. Distribution of amphipathic n-grams in different proteomes

As mentioned before, the statistical test for all the analyzed sequences does reject the hypothesis of uniformity (see Appendix). Second, the non-uniformity is readily seen in the respective histograms: the amphipathic n-grams occur more often amongst upper ranks. The distributions for individual organisms, though, are different in details.

The graph of the distribution of amphipathic 12-grams (Figure 1) shows, that there is always more amphipathic 12-grams at the upper part of the sorted vocabulary (left half of the histogram), so that in the distributions all have general descending character. For seven of 10 species, *Arabidopsis thaliana, Bos taurus, Dictyostelium discoideum, Homo sapiens,*

Micromonas sp., Plasmodium yoelii, and *Saccharomyces cerevisce,* the first two bins show lower values, compared to other species (data not shown). The trend, however, is general descent (bins 3–10). Thus, the amphipathic 12-grams are typically among the frequent words than would be expected on random basis. For the proteomes of *Bacillus subtilis, Trypanosoma brucei,* and *Monosiga brevicolis* the curves monotonically decrease along the whole range. The differences between investigated organisms can be caused by several reasons. The amphipathicity sequence code is definitely not the only code present in proteome. There are other codes in protein sequences and various groups of organisms use them in different ways. The usage of the amphipathic vocabulary could also diversify during evolution.

From the whole data-set the histograms for two organisms showing opposite extreme deviations at the

initial bins are shown (*Trypanosoma brucei* – thin continuous line, *Plasmodium yoelii* – dotted line). The thick line corresponds to the average for 10 species listed in the Table 2.

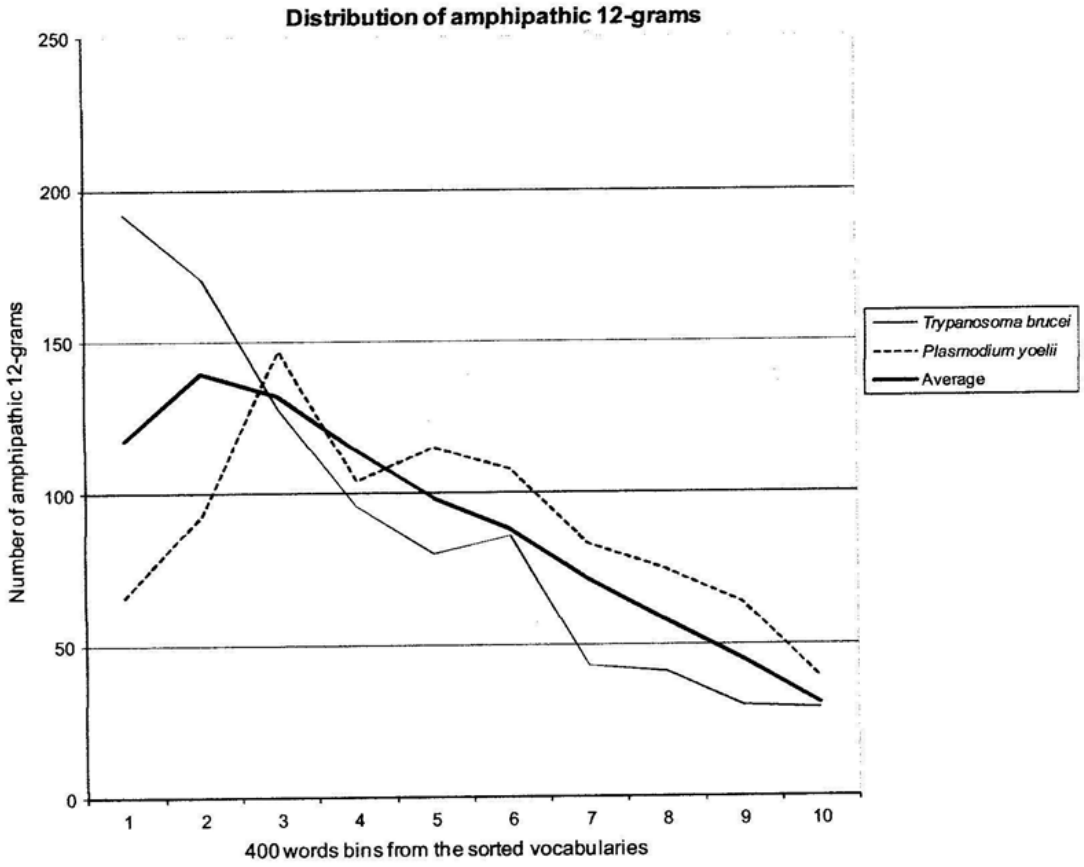


Figure 1. The distribution of amphipathic 12-grams in the list of total vocabulary divided into 10 bins.

3.2. Distribution of consonants and vowels in three Indo-European languages

To analyze precisely the ratio of vowels and consonants in different languages, a rigorous way would be to transform investigated texts to their phonetic transcription according to some uniform algorithm. Such an algorithm is a problem *per se* as it should respect all phonetic specifics of the languages which we consider excessive for our half-quantitative study. Our aim is only to show analogy between genetic text and human text, without exact quantification of mathematical properties of the texts. From the lists of n-grams in Table 3, it is seen that the most frequent n-grams are strings of alternating C (CC) and V (VV). (Contribution of alternating words in humantexts analyzed according to this model for n-grams of length 4, 5, and 6 varies between 57 and 89%.)

The lists contain n-grams of length 4, 5, and 6 because it is a range where most words usually occur. As the analyzed texts have been fused in long sequences, without blanks, most frequent combinations do not always correspond to existing words but also to connected parts of neighboring words, such as EXISTING WORDS (VCCCVC).

Alternating n-grams are shown in capital letters. In all cases, the alternating pronounceable combinations are at the top. The non-alternating words with more than two consonants or vowels in a row (lower type) appear more towards the bottoms of the lists. The contribution of alternating words in the upper parts of the ranked lists range between 90 and 100%.

4. Discussion

The aim of this comparison is to illustrate one expected analogy between human texts and protein sequences. There are natural constraints which produce specific types of alternation in both types of the texts. Distribution of amphipathic words in genetic texts seems to be species specific as anomalies at the start of the histograms in the Figure 1 indicate. This may serve as kind of “proteomic signature” for potential further exploration. The relatively smaller proportion of amphipathic words at the very tops of the sorted lists may be explained by strong domination of the N-rich and P-rich peptides in eukaryotic proteomes. The P-rich

peptides would correspond to so-called naturally unfolded (intrinsically disordered) proteins (Dunker et al., 2000).

The hydrophobicity, similar to syllabicity, is not purely discrete feature, and some hydrophobic residues (such as A and G) sometimes behave in context of adjacent amino acids neutrally. In similar way, e.g. in Slavic languages, there are some consonants which can behave as vowels under certain conditions when they stand in the place of vowels and they actually produce syllables (such as L, R). These exceptions do not change the general observation on the dominance of the alternates. The type of analysis described can not be fully applied to real phonetic situation, which is too complicated and can even vary not only in dialects of the same language but also between different speakers. The pronounceability, however, remains a demand of every language, irrespective of its phonetic or written presentation.

Any detailed analogy between genetic texts and human languages could be pursued only very cautiously as, indeed; these are principally different "scripts". Those comparisons that can be and have been made, however, have a merit of seeing each of the systems from a new perspective, and bringing, therefore, some new understanding.

Table 3. Extreme ranks (top and bottom) from the sorted n-grams (n=4, 5, 6).

English	Frequency	French	Frequency	Czech	Frequency
4-grams					
CCVC	2700	VCCV	2317	VCVC	3121
CVCC	2621	CVCV	2310	CVCV	3112
CVCV	2180	VCVC	2243	CVCC	2060
VCVC	2102	CCVC	2131	CCVC	2050
VCCV	2041	CVCC	2063	VCCV	1697
VVCV	367	vccc	613	CCVV	170
ccccc	266	cvvv	228	VVCC	161
vvvc	83	vvvc	228	cccc	107
cvvv	83	cccc	45	vvvc	18
cvvv	15	vvvv	27	cvvv	18
5-grams					
CVCVC	1783	CVCVC	1675	CVCVC	2923
VCCVC	1704	VCCVC	1671	VCVCV	1903
CVCCV	1689	CVCCV	1597	CVCCV	1563
CCVCC	1436	VCVCV	1185	VCCVC	1553
CCVCV	1264	CCVCV	1125	VCVCC	1218
ccccc	31	vcccc	44	ccvvv	9

vvvcv	17	cvvvv	24	vcvvv	9
vvvvc	14	vvvvc	24	vvvcc	8
cvvvv	14	vvvvv	3	ccccc	8

vvvvv 1 ccccc 1 VVCVV 7

6-grams

CVCCVC	1420	CVCCVC	1161	VCVCVC	1795
CCVCVC	1033	VCCVCV	894	CVCVCV	1782
CVCVCC	999	CVCVCV	875	CVCCVC	1447
CCVCCV	950	VCVCVC	839	CVCVCC	1141
VCCVCC	907	CCVCVC	836	CCVCVC	1127
vvvcvv	4	vvcccc	6	vccccc	8
cccccc	2	cvvvvv	3	CVVCVV	7
cvvvvv	1	vvvvvc	3	VVCVVC	7
vvvvcv	1	vccccc	1	ccccvv	4
vvvvvc	1	cccccv	1	vvcccc	1

Acknowledgments

Discussion with A.E. Rapoport is highly appreciated. The work has been supported by the Czech Ministry of Education (grant MSM0021622415 and MSM0021620828) and by the project

"Innovation of the General linguistics and Theory of communication in cooperation with the natural sciences" (grantCZ.1.07/2.2.00/28.0076).

References

Beckmann, J. S., Brendel, V., & Trifonov, E. N. (1986). Intervening sequences exhibit distinct vocabulary. Journal of Biomolecular Structure Dynamics, 4, 391–400. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2482751

Berezovsky, I. N., Grosberg, A. Y., & Trifonov, E. N. (2000). Closed loops of nearly standard size: Common basic element of protein structure. FEBS Letters, 466, 283–286. doi:10.1016/S0014-5793(00)01091-7

Bolshoy, A. (2003). DNA sequence linguistic tools: Contrast vocabularies, compositional spectra and linguistic complexity. Applied Bioinformatics, 2, 103–112. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15130826

Brendel, V., Beckmann, J. S., & Trifonov, E. N. (1986). Linguistic of nucleotide sequences: Morphology and comparison of vocabularies. Journal of Biomolecular Structure Dynamic, 4, 11–21. Retrieved from http://www.ncbi.nlm. nih.gov/pubmed/3078230

Clark, J., & Yallop, C. (1995). An Introduction to phonetics and phonology (2nd ed.). Oxford: Blackwell.

Dunker, A. K., Obradovic, Z., Romeo, P., Garner, E. C.,Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. Genome Informatics, 11, 161–171. Retrieved from http://www.jsbi.org/pdfs/journal1/GIW00/GIW00F16. Pdf

Eco, U. (1976). Theory of semiotics. Bloomington, IN: Indiana University Press.

Epand, R. M. (Ed.). (1993). The amphipathic helix. Boca Raton, FL: CRC Press.

Fink, A. L. (2005). Natively unfolded proteins. Current Opinion in Structural Biology, 15, 35–41. Retrieved from http://www.sciencedirect.com/science/article/pii/S0959440X05000035

Ganapathiraju, M., Weisser, D., Klein-Seetharaman, J., Rosenfeld, R., Carbonell, J., & Reddy, R. (2002). Comparative n-gram analysis of whole-genome sequences. Human Language Technologies Conference, San Diego. Retrieved from http://dl.acm.org/citation.cfm?id=1289259 & dl=ACM & coll=DL & CFID=188588611 & CFTOKEN=86531065

Hall, Ch. J. (2005). An introduction to language and linguistics: Breaking the language spell (pp. 110–130). London: Continuum.

Hausman, R. E., & Cooper, G. M. (2004). The cell: A molecular approach. Washington, DC: ASM Press.

Lehmann, W. P. (1992). Historical linguistics: An introduction (3rd ed., pp. 1–22). London: Routledge Matisoff.

Mant, C. T., Zhou, N. E., & Hodges, R. S. (1993). The role of amphipathic helices in stabilizing peptide and protein structure. In R. M. Epand (Ed.), The amphipathic helix (pp. 38–54). Boca Raton, FL: CRC Press.

Osmanbeyoglu, H. U., & Ganapathiraju, M. K. (2011). N-gram analysis of 970 microbial organisms reveals presence of biological language models. BMC Bioinformatics, 12, 1–12. doi:10.1186/1471-2105-12-12

Pietrokovski, S., & Trifonov, E. N. (1992). Imported sequences in mitochondrial yeast genome identified by nucleotide linguistics. Journal of Biomolecular Structure Dynamics, 7,1251–1268. Retrieved from http://www.sciencedirect.com/science/article/pii/037811199290040V

Popov, O., Segal, D. M., & Trifonov, E. N. (1996). Linguistic complexity of protein sequences as compared to texts of human languages. Biosystems, 38, 65–74. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8833749

Rapoport, A. E., & Trifonov, E. N. (2011). “Anticipated” nucleosome positioning pattern in prokaryotes. Gene, 488, 1–2. doi:10.1016/j.gene.2011.08.002

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379–423, 623–656.

Sheskin, D. J. (2004). Handbook of parametric and nonparametric staistical procedures (3rd ed., pp. 203–217). Washington, DC: Chapman and Hall/CRC.

Trifonov, E. N. (1989). Viewpoint: Multiple codes of nucleotide sequences. Bulletin of Mathematical Biology, 51, 417–432. Retrieved from http://link.springer.com/article/10.1007%2FBF02460081?LI=true#page-1

Volkovich, Z., Kirzhner, V., Bolshoy, A., Nevo, E., & Korol, A. (2005). The method of n-grams in large-scale clustering of DNA texts. Pattern Recognition, 38, 1902–1912. doi:10.1016/j.patcog.2005.05.002

Appendix

Table A1. Kolmogorov–Smirnov statistical test.

Organism	Test statistic D	Critical value D_n (0.05)
Bacillus subtilis	0.266	0.021
Plasmodium yoelii	0.127	0.021
Trypanosoma brucei	0.295	0.021
Monosiga brevicolis	0.174	0.021
Dictyostelium discoideum	0.141	0.021
Saccharomyces cerevisiae	0.191	0.021
Micromonas sp.	0.170	0.021
Arabidopsis thaliana	0.219	0.021
Bos taurus	0.201	0.021
Homo sapiens	0.190	0.021

Notes: In all cases test statistic D (at the level of significance 0.05) exceeds critic value D_n (0.05). Hence the hypothesis of uniformity of distribution of amphipathic 12-grams is rejected. The randomly selected samples from proteomes are of the uniform length of 1 000 000 of 12-grams, thus the critical values are equal.

3. Zahradník D, Trifonov EN, Zemková M. (2015) Evolutionary landscape of human genome vocabulary. Czech and Slovak linguistic review

Abstract

Inspection of a full vocabulary of the words (16-mers) of human genome reveals that the top of the list ranked by word occurrence contains almost exclusively simple repeats, and words from Alu sequences – most abundant dispersed elements. These excessive words can be considered as “generators” and suggest a simple model of genome evolution: everlasting intrusion of the generator sequences in “neutral” regions of the genome, and gradual mutational changes, causing the increase in the sequence complexity. The way to detect the generators is to find those of which point mutated forms appear less frequently. Examples of the generators are presented**.**

Keywords

Evolution of genomic sequences; sequence generators; genome vocabulary

Introduction

Sequences of bases of DNA, sometimes of many millions of letters, may be considered as continuously written text of the genomic language in which "sentences" and "paragraphs" can be recognized by their specific sequence markers, short "words". The continuous writing is known well in some ancient scripts, e.g., Etruscan. Some types of the nucleotide sentences may be further subdivided in various kinds of words. The triplets of the protein coding sequences are one example. The volume of information carried by the sequences is huge, very much due also to the overlapping of various messages carried by the sequences – feature not known in human texts. The subdivision of the sequences in the words and sentences is very far from completion, and the actual biological meanings of those recognized sequence segments are largely unknown. This is why the sequences often are split in formal *k*-tuples – continuous stretches of *k* letters not necessarily corresponding to specific functional words. Counting all the *k*-tuples of various sizes (with overlapping) results in the frequency "vocabularies" of the sequences, which are important characteristics of the sequences, useful for elucidation of possible functional involvements of the *k*-tuples (called words further on) and of evolutionary scenarios for the genomes. Such are vocabularies

computed from human genome sequences – an important source for their functionally and evolutionarily related analyses. Studying the vocabulary of genomic sequences is analogous to the research on human language vocabularies, both highly related to evolution of respective scripts.

One aspect of the analysis of the vocabularies – an attempt to detect remnants of the most ancient sequences, those from which the genomes started their evolution. If, for example, the ancient genomes evolved by genome duplication (resulting in twice larger genome size) as, it is believed, the modern genomes still do (Ohno, 1971), then the respective words of the vocabulary would duplicate as well, gradually increasing their numbers in the course of evolution, on the background of concomitant mutational changes. The surviving words, in multiple copies, unchanged since the earliest times, thus, would represent the ancient sequences. If the sequences appeared again and again in evolution, they should have elevated copy numbers in the genomes today. One can imagine, of course, that some more recent invaders would have generated more words than the number of remnants of the ancestor.

Materials and methods

For illustrational purposes we used human genome sequences which were downloaded from NCBI database (National Centre for Biotechnology Information, 2014). (However any other eukaryotic genome could be used). To simplify calculation we took for demonstration of sequence generators a random sample of 10 Mb. The selection was done in two steps: First, 10 000 genes from the genome were randomly chosen. Secondly, from each such a gene a 1000 b long sequence was randomly selected.

Than the list of 15- mers was generated and sorted according to frequency. For several topmost sequences we found their “offsprings” (lists of all one-point mutations of original sequence).

For all computations we used our own software tool “Complexity_ G” available at http://web.natur.cuni.cz/filosof/index.php/cs/clenove/362-zemkova.html

Results and discussion

Human genome vocabulary

In the Fig. 1 there is presented a small part of the human genome vocabulary of 15-mers (Rapoport & Trifonov, 2013). The most frequent words nos. 1–6, 9 and 10 are all simple tandemly repeating sequences. In random sequence of human genome size any such word would appear just few times, if at all. What is observed, however, exceeds by orders of magnitude, the expected random occurrences. This also means that there is a substantial number of theoretically possible words which do not appear at all, which makes these missing words of special interest (suicidal letter combinations? "dirty words"?).

1	1198780	TTTTTTTTTTTTTTT **(T)n**
2	1190667	AAAAAAAAAAAAAAA **(A)n**
3	366285	TGTGTGTGTGTGTGT **(TG)n**
4	362623	ACACACACACACACA **(AC)n**
5	348215	GTGTGTGTGTGTGTG (TG)n
6	344421	CACACACACACACAC (AC)n
7	**223424**	**GCTGGGATTACAGGC Alu**
8	**223011**	**GCCTGTAATCCCAGC Alu**
9	222894	TATATATATATATAT **(TA)n**
10	222730	ATATATATATATATA (TA)n
Alu		
68	169033	TTTTTTTTTTTTTT**G**
Alu		
73	167889	**C**AAAAAAAAAAAAAA
Alu		
75	150349	**C**TTTTTTTTTTTTTT
76	149748	AAAAAAAAAAAAAA**G**
Alu		
83	138448	TTTTTTTTTTTTT**GA**
84	137643	**TC**AAAAAAAAAAAAA
85	135070	TTTTTTTTTTTT**GAG**
86	134465	TTTTTTTTTTT**GAGA**
87	134262	**CTC**AAAAAAAAAAA
88	133917	**TCTC**AAAAAAAAAA

Fig. 1. Frequency vocabulary of 15-mers of human genome, the topmost words. "Missing" lines labelled "Alu" correspond to various segments of the generic 300 bases long sequence.

Hundreds of thousands of simple repeating words in the human genome is result of their "expansion" during DNA replication – generation of extra copies of elementary repeating units (mono- and dinucleotides in this case) by the slippage mechanism (Wells, 1996), Having been generated this way, the repeating sequences subsequently gradually change, mutate, replacing the monotonous elements by other letters here and there. The recent paper (Frenkel & Trifonov, 2012) outlines and substantiate the theory of origin and evolution of genomes whereby such expanding simple sequences brought the genomes into being and continued to both expand and mutate, so that today we detect in the genome sequences the generic simple sequences, their recognizable mutated versions (see nos. 68, 76, 83–88 in the Fig. 1), and the words of higher complexity such that their original simplicity is severely compromised by mutations and not recognized anymore. One handy example is so-called Alu sequence, a frequent complex sequence repeat of length ~300 bases scattered all over genome in over 200 thousands of mostly isolated (non-tandem) copies (Batzer & Deininger, 2002).

The Alu sequence has its own long evolutionary history. One could think of, say, the $(\mathbf{GGT})_n$ repeat as ancient precursor of Alu (compare with No. 7 in Fig. 1: **G**C**TGG**GAT**T**ACA**GG**C, the most frequent segment of Alu). Today this sequence, diverged very far from the ancestor, continues to invade the human genome, and mutate further (see below).

Sequence generators

One may ask a fundamental question: What was the very first, perhaps primitive, nucleotide sequence from which genomes started their spectacular evolution towards *Homo sapiens*? Probably, this generic sequence existed in many copies and continued to invade the early genome(s), by expansion (for simple repeats), or by transposition (like more complex Alu sequences). These invading sequences, appearing in many copies, generated the genomes, originally very simple but diversified further towards higher complexity. We, then, may ask a different question: what are the sequence generators today (hoping that among them is also *the* original generator).

Conservatively, one would suggest that all high occurrences in the vocabulary are due to higher use of the words for whatever special intragenomic function they serve. The most conspicuous of the high scores are those which belong to simple repeats or transposable

elements. Many of the simple repeats have specific functions already by their presence often within protein-coding sequences. Another well documented function is modulatory one (Trifonov, 1989, 2004). The aggressive behaviour of the tandem repeats, their invasive character (2/3 of human genome are the repeats) suggest one more, very general role for the repeats – as source of new sequences for genome evolution. In this sense the high scores in the vocabulary would represent sequence generators. The whole theory describing the genome as product of activities of such generators has been recently suggested (Frenkel & Trifonov, 2012). Thus, the reasons for the formation of the high occurrence words are either their higher functional use or their generic function (generators), or both. Mutations of the high occurrence words would bring the scores down, while fresh emergence of the generator words in the genome would increase the scores. Modest scores, which are still higher than for all respective point mutations of the word, may mean that the corresponding generator is not as active anymore as some time before, or this is not generator at all, but one of functional favourite words of the genome. The sharper is the generator (more numerous than all its point mutated versions), the higher is its role in evolution as supplier of material for mutated versions with eventual functionality and higher use.

In the Fig. 2 the family of point mutations of the generator $(\mathrm{A})_{16}$ is presented. Interestingly, during the accumulation of the point changes the original pool of 3122 AAAAAAAAAAAAAAAA has been exhausted, but new 3254 AAAAAAAAAAAAAAAA have been generated in the mean time. Similar result of comparable numbers of the Alu-generator TAATCCCAGCACTTTG and its mutated family is shown in the Fig. 3. The fact that the number of copies of the generators is comparable to the number of their point mutation derivatives suggests that the two processes of emergence of new copies of generator, and of mutations in the generator sequence are in a steady state. Other examples of the generators and their families are presented in the Figure 4.

AAAAAAAAAAAAAAAA		
3254	AAAAAAAAAAAAA**C**AA	AAAA**T**AAAAAAAAAAA
CAAAAAAAAAAAAAAA	40	12
525	A**G**AAAAAAAAAAAAAA	AAAAAA**C**AAAAAAAAA
AAAAAAAAAAAAAAA**G**	40	11
469	AAAAA**G**AAAAAAAAAA	AA**T**AAAAAAAAAAAAA
AAAAAAAAAAAAAA**G**A	39	11
284	AAAAAAAAAAAA**C**AAA	AAA**T**AAAAAAAAAAAA
AAAAAAAAAAAAA**G**AA	36	10
219	AAAAAAAA**G**AAAAAAA	AAAA**C**AAAAAAAAAA
AAAAAAAAAAAAAAA**T**	35	9
196	AAAA**G**AAAAAAAAAA	AAAAAAAAAA**T**AAAAA
AAAAAAAAAAAA**G**AAA	34	8
166	AAAAAAA**G**AAAAAAAA	AAAAAAAAA**T**AAAAAA
TAAAAAAAAAAAAAAA	33	8
145	AAA**G**AAAAAAAAAAAA	AAAAA**T**AAAAAAAAAA
AAAAAAAAAA**G**AAAA	29	8
104	AAAAAAAAAA**C**AAAA	AAA**C**AAAAAAAAAAAA
AAAAAAAAAAAAAAA**C**	28	7
78	AAAAAAAAA**C**AAAAA	AA**C**AAAAAAAAAAAAA
GAAAAAAAAAAAAAAA	27	7
73	AAAAAAAA**C**AAAAAA	AAAAAAAAAAAAA**T**AA
AAAAAAAAA**G**AAAAA	26	7
59	AA**G**AAAAAAAAAAAA	AAAAAA**T**AAAAAAAAA
A**C**AAAAAAAAAAAAAA	23	7
52	AAAAAAA**C**AAAAAAA	AAAAAAAA**T**AAAAAAA
AAAAAA**G**AAAAAAAAA	17	6
45	A**T**AAAAAAAAAAAAAA	AAAAAAAAAA**T**AAAA
AAAAAAAA**G**AAAAAA	16	6
45	AAAAAAA**C**AAAAAAAA	AAAAAAAAAAA**T**AAA
AAAAAAAAAAAAAA**C**A	14	5
44	AAAAA**C**AAAAAAAAA	AAAAAAA**T**AAAAAAAA
AAAAAAAAAAAAAA**T**A	12	5
42		sum 3122

Fig. 2. The tandem generator $(A)_{16}$ and "family" of its point mutations (bold). The calculation is made on 10 Mbase subset of human genome, listing all words of length 16 letters.

TAATCCCAGCACTTTG	TAATCCCAGCA**A**TTTG	TAATCCCAGCACTT**G**G
800	13	6
TAATCCCA**A**CACTTTG	TAATC**A**CAGCACTTTG	TAATCCCAG**G**ACTTTG
64	12	6
TAATCC**T**AGCACTTTG	TAATCCC**G**GCACTTTG	TAATCCCAGC**T**CTTTG
59	11	6
TAATCCCAGCA**T**TTTG	TAATCCCAGCAGTTTG	TAATCCCAGCACT**A**TG
37	10	6
TAATC**T**CAGCACTTTG	TAATCCCAGC**C**CTTTG	TAATCCCAGCAC**C**TTG
36	9	6
TAAT**T**CCAGCACTTTG	TAATCCC**T**GCACTTTG	TAATCC**A**AGCACTTTG
34	9	6
TAATCCCAG**T**ACTTTG	TAAT**G**CCAGCACTTTG	TAATCCCA**T**CACTTTG
33	9	6
TAATCCCAGCACT**C**TG	T**T**ATCCCAGCACTTTG	TAAT**A**CCAGCACTTTG
21	9	6
TAATCCCAGCACTTT**A**	TAATCCCAGCACTTT**C**	TAATCCCA**C**CACTTTG
20	8	5
TAATCCCAG**A**ACTTTG	TA**G**TCCCAGCACTTTG	TAATC**G**CAGCACTTTG
18	8	5
TAATCCCAGCACTTT**T**	TAATCCCAGC**G**CTTTG	**A**AATCCCAGCACTTTG
15	7	5
TAATCCCAGCACTT**C**G	**G**AATCCCAGCACTTTG	TA**C**TCCCAGCACTTTG
15	7	3
CAATCCCAGCACTTTG	TAATCCCAGCACT**G**TG	TAATCCCAGCAC**A**TTG
15	7	3
TAA**C**CCCAGCACTTTG	TAATCCCAGCAC**G**TTG	TAA**A**CCCAGCACTTTG
14	7	3
T**C**ATCCCAGCACTTTG	TAATCCCAGCACTT**A**G	TAATCC**G**AGCACTTTG
14	6	2
T**G**ATCCCAGCACTTTG		sum
13		614

Fig. 3. The non-tandem Alu generator TAATCCCAGCACTTTG and "family" of its point mutations (bold). The calculation is made on 10 Mbase subset of human genome, listing all words of length 16 letters. Note that in this sequence ensemble the topmost Alu word is different compared to the Figure 1 data.

TGTGTGTGTGTGTGTG	TGTGTGTGT**C**TGTGTG	TGTGTGTGTG**C**GTGTG
1200	19	10
TGTGTGTGTGTGTGT**A**	T**C**TGTGTGTGTGTGTG	TGTGTGTGTGTG**C**GTG
74	18	9
CGTGTGTGTGTGTGTG	TGTGTGT**C**TGTGTGTG	TGTGTGTGT**T**TGTGTG
35	17	8
TGTGTGTGTGTGT**A**TG	TGTGTG**C**GTGTGTGTG	TGTGTGTGTG**A**GTGTG
32	16	7
TGTGTGTGTGTGTG**A**G	TGTGTGTGTGTGT**T**TG	TGTGTGTG**A**GTGTGTG
32	15	7
TGTGTGTGTGTGTGT**T**	TGTGTGTGTGT**C**TGTG	TGTGTGTGTGTG**A**GTG
32	15	7
AGTGTGTGTGTGTGTG	TGTGTGTG**C**GTGTGTG	TGTG**A**GTGTGTGTGTG
32	14	6
TGTGT**A**TGTGTGTGTG	TGTGTGTGTGTGTGT**C**	TGTGTG**A**GTGTGTGTG
29	14	5
GGTGTGTGTGTGTGTG	TGTGTGTGTGT**T**TGTG	TGTGTGT**T**TGTGTGTG
28	12	4
T**A**TGTGTGTGTGTGTG	TGTGT**C**TGTGTGTGTG	TGTGT**T**TGTGTGTGTG
26	12	4
TGTGTGTGTGT**A**TGTG	TGTGTGTGTGTGT**C**TG	TGTG**G**GTGTGTGTGTG
25	12	4
T**T**TGTGTGTGTGTGTG	TGTG**C**GTGTGTGTGTG	TGTGTG**G**GTGTGTGTG
25	12	2
TGTGTGT**A**TGTGTGTG	TG**A**GTGTGTGTGTGTG	TGT**T**TGTGTGTGTGTG
24	12	2
TGTGTGTGT**A**TGTGTG	TG**C**GTGTGTGTGTGTG	TG**G**GTGTGTGTGTGTG
24	11	1
TGTGTGTGTGTGTG**C**G	TGT**C**TGTGTGTGTGTG	TGTGTGTGTGTG**G**GTG
21	10	1
TGT**A**TGTGTGTGTGTG		sum
20		745

TATATATATATATATA	TATATATATATATA**A**A	T**C**TATATATATATATA
321	15	6
TATATATATATATA**C**A	TATATA**C**ATATATATA	TAT**T**TATATATATATA
40	14	6
AATATATATATATATA	TATATATA**A**ATATATA	TATATATATATAT**T**TA
31	14	6
T**G**TATATATATATATA	TATATATATATA**C**ATA	TATATAT**C**TATATATA
31	14	4

CATATATATATATATA	TATATATATAT**G**TATA	TATATATATA**A**ATATA
31	13	4
TATATATATATATAT**G**	TATA**C**ATATATATATA	TATATATAT**C**TATATA
26	12	4
TATATATATATATAT**T**	TATATATATATATAT**C**	TATATATATAT**T**TATA
25	11	4
TATATAT**G**TATATATA	TATATA**A**ATATATATA	TATATATATATATA**G**A
22	11	4
T**T**TATATATATATATA	TAT**G**TATATATATATA	TATATATATATAT**C**TA
20	11	3
TATAT**G**TATATATATA	TATATAT**T**TATATATA	TATAT**T**TATATATATA
18	9	2
TATATATATA**C**ATATA	**G**ATATATATATATATA	TATAT**C**TATATATATA
17	9	2
TATATATAT**G**TATATA	TATA**A**ATATATATATA	TATATATATATA**G**ATA
17	9	1
TATATATA**C**ATATATA	TA**A**ATATATATATATA	TATATATATAT**C**TATA
16	8	1
TA**C**ATATATATATATA	TATATATAT**T**TATATA	TATATATATATA**A**ATA
15	7	1
TATATATATATAT**G**TA		sum
15		529

GAGAGAGAGAGAGAGA		
184	GAGAGAGAGAGAGAG**T**	GAGAGAGAGAGAG**G**GA
GAGAGAGAGAGAGA**A**A	8	3
19	GAGAGAGA**C**AGAGAGA	GAGAGA**T**AGAGAGAGA
GAGAGAGAGAGA**A**AGA	7	3
17	GAGAGAGA**A**AGAGAGA	GAGA**T**AGAGAGAGAGA
CAGAGAGAGAGAGAGA	6	3
15	GAGAGA**C**AGAGAGAGA	GAGAG**G**GAGAGAGAGA
GA**A**AGAGAGAGAGAGA	6	2
13	GAGAGA**A**AGAGAGAGA	GAG**G**GAGAGAGAGAGA
AAGAGAGAGAGAGAGA	6	2
12	GAGA**A**AGAGAGAGAGA	GAGAGAGAGAGAG**T**GA
GAGAGAGAGA**A**AGAGA	6	2
11	**T**AGAGAGAGAGAGAGA	GA**T**AGAGAGAGAGAGA
GAGA**C**AGAGAGAGAGA	5	2
11	GAGAGAGAGA**C**AGAGA	GAGAGAGAGAGAGAG**C**
GA**C**AGAGAGAGAGAGA	5	1
10	GAGAGAGAG**G**GAGAGA	GAGAGAGAG**T**GAGAGA
GAGAGAGAGAGAGA**C**A	5	1

10	G**G**GAGAGAGAGAGAGA	GAGAG**C**GAGAGAGAGA
G**T**GAGAGAGAGAGAGA	4	1
10	GAGAGAG**G**GAGAGAGA	GAG**T**GAGAGAGAGAGA
GAGAGAGAGAGA**C**AGA	4	1
8	GAGAGAGAGAG**G**GAGA	GAGAGAG**T**GAGAGAGA
GAGAGAGAGAGAGAG**G**	4	1
8		sum
		232

GCCGCCGCCGCCGCCG		
12	GCCGCCGCCGCC**A**CCG	GCCGCC**C**CCGCCGCCG
GC**T**GCCGCCGCCGCCG	1	1
2	GCCG**G**CGCCGCCGCCG	GCC**C**CCGCCGCCGCCG
GCCGCCGCCGCCGCC**C**	1	1
2	GCCGC**T**GCCGCCGCCG	GCCGCCGCCGCCGCC**A**
GCCGCCGCCGCCGC**T**G	1	1
2	GCCGCCGCCGC**A**GCCG	G**A**CGCCGCCGCCGCCG
CCCGCCGCCGCCGCCG	1	1
2	GCCGCCGCC**A**CCGCCG	**A**CCGCCGCCGCCGCCG
GCCGCCGCCGCCGCC**T**	1	1
2	GCCGCC**A**CCGCCGCCG	sum 22
GCCGCCGCCGCC**T**CCG	1	
1		

Fig. 4. The generators TGTGTGTGTGTGTGTG, TATATATATATATATA, GAGAGAGAGAGAGAGA and GCCGCCGCCGCCGCCG, and the "families" of their point mutations (bold). The calculation is made on 10 Mbase subset of human genome, listing all words of length 16 letters.

Some properties of the generators

Generator words emerge by expansion of repeats, by transposition from outside, or by multiplication inside. They appear only in neutral sites of the genome, as illustrated, for example, by phenomenon of nesting of transposons (SanMiguel et al. 1996, 1998), and they, essentially, do not interfere with genome activities. Their mutations lead either to lethality, or to survival staying neutral, or to some positive effect (functionality), in which case the corresponding mutated word may accumulate. Thus, the generator will be accompanied by 3*n* (for *n*-letter words) or less relatives of various occurrences. In the next rounds of relatives

(two and more point changes) the weights of some of the nodes may further increase. The generator power is measured by its relative amplitude – ratio of its weight to average occurrence of all its point-mutated variants. In the ranking of generators by power, the highest ones would be the likeliest generators.

The non-renewable generators are expected to slowly decrease in evolution. The topmost of them are fresh newcomers, like Alu sequences.

From any node one could come to any other node by *n* or less steps, in many alternative ways. To find the likeliest ancestor, one has to choose the trajectory of minimal length to the nearest generator. One simple strategy is to take consecutively the highest neighbour (steepest ascent). The node may belong to "island", so that the generator proper may not be reached (lost intermediate ancestry). The number of such nodes is expected to be small. The detection of the likely ancestor, however, is not the proof of the ancestry. At best, one can expect the consistency of the results with the simple model of mutating generators.

Since there is no known mechanism of acquiring the same transposable element again and again in evolution (they evolve themselves, outside of the genome), the fate of such generators is their gradual decay ending with modest vocabulary weights. It would be, probably, hard to find the ancestor of this node, as it came, most likely, from unrelated vocabulary.

Conclusion

The exceptionally high numbers of the simple repeat words in the genome vocabulary suggest that they served as generators a long time, since the mechanism of their formation (slippage during replication) exists from the moment of the emergence of the sequences. Such sequences, therefore, may be considered both as the original sequences of primitive genomes, and as building material of the genomes in the course of their further evolution, from which more and more sophisticated sequences eventually formed. The study is the first step in an overview of what would be the original and more modern genome-forming sequence generators.

In further studies, other genomes have to be analyzed as well, but this would be subject of the whole new field of "nucleotide linguistic research" - about origin of genes and genomes, which has been initiated only very recently (Frenkel and Trifonov, 2012).

References

Batzer, M. A. – Deininger, P. L. (2002): Alu repeats and human genomic diversity. Nature Reviews 3, p. 370–380.

Frenkel, Z. M. - Trifonov, E. N. (2012): Origin and evolution of genes and genomes. Crucial role of triplet expansions. J Biomol Struct Dyn, 30(2), p. 201–210.

Rapoport A. E., Trifonov E. N. (2013): Compensatory nature of Chargaff's second parity rule. J Biomol Struct Dyn, 31(11), 1324–1336.

Trifonov, E. N. (1989): The multiple codes of nucleotide sequences. Bull Math Biol, 51(4), p. 417–432.

Trifonov, E. N. (2004): Tuning function of tandemly repeating sequences: a molecular device for fast adaptation. Evolutionary theory and processes: Modern horizons, Springer: p. 115–138.

Wells, R. D. (1996): Molecular basis of genetic instability of triplet repeats. Journal of Biological Chemistry, 271(6), p. 2875–2878.

Ohno, S. (1970): *Evolution by Gene Duplication.* New York: Springer-Verlag.

SanMiguel, P. Tikhonov, A. Jin, Y. K. Motoulskaia, N. Zakharov, D. Melake-Berhan, A. Springer, P. S. Edwards, K. J. Lee, M. Avramova, Z. et al. (1996): Nested retrotransposons in the intergenic regions of the maize genome. Science 274, 765–768.

SanMiguel, P. Gaut, B. S. Tikhonov,A. Nakajima, Y. Bennetzen, J. L. (1998): The paleontology of intergene retrotransposons of maize. Nat Genet 20, 43–45.

Internet data source:

National Centre for Biotechnology Information: http://www.ncbi.nlm.nih.gov/ (Accessed 20 September 2014)

4. Trifonov EN, Zemková M. (2015) Genome and language – two scripts of heredity (Ontogenetic theory of language origin). Czech and Slovak linguistic review

Glossary

Since the paper is of an interdisciplinary character, we introduce here major terms and their commonsensical meanings in the context of the paper, to minimize misunderstanding. For more rigorous definitions see respective dictionaries.

Language A system of signals (vocal, gestural or written) used for communication between individuals

Sequence Nucleotide sequence of DNA or RNA, and amino acid sequence of proteins – text of nucleotide or amino acid symbols. Similarly – sequences of letters in human texts. For example, the baby babbling is sequence of incoherent uttering.

Script System of organized symbols to convey information in form of text

Symbol Element of sequence or text (letter, corresponding to phoneme, or nucleotide, or amino acid). In more general sense – also pictograms and syllables.

Text Linear array of letters. Written representation of any language.

Letter Written notation for phoneme or chemical symbols (amino acids, nucleobases)

Glottogenesis - discipline studying origin of language and speech

Abstract

One manifestation of life which is not easy to comprehend is repeated invention of linear script, first as genomic sequences, and then – as language writings. Both possess their specific alphabets. Evolution of both started, likely, from simple repetitions (TGTGTG…, GCCGCCGCC…, ma-ma-ma, da-da-da). The ontogenetic dimension comes from the fact that the spectrum of the easiest pronounced simple repeats by babies in the period of canonical babbling is the same for all ethnicities, and, quite likely, it represents the early hominids` vocal abilities. Later the repetitions continued to enter, but simultaneously they accumulated mutational changes, turning in more complex words already not recognizable as repeats. This scenario appears to be common for both- genome and language, which, thus, carry the heritage of the human species each in its way.

Key words: evolution of genomic sequences; evolution of language; canonical babbling; triplet expansion; ontogeny of speech

Introduction

Genome and language – two scripts of heredity

Enigmatic and tantalizing is the fact that both genetic texts (nucleic acid and protein sequences) and human scripts can be written in the form of linear strings of symbols. Both carry various kinds of information, maintained in generations either by DNA replication (for genetic sequences), or by rewriting and reprinting (language texts). One can, of course, safely call it an analogy, though the connection appears to be much deeper. Phenomenon of life as generator of genetic information has evolved on the basis of the sequences in the epoch of species formation, and on the same basis of linear script the life entered in its most advanced epoch - formation of human culture and languages.

The evolutionary nature of both genetic and language scripts suggests that, perhaps, there are some similarities in their origin and evolution. In particular, the nucleotide sequences are believed to evolve from simple tandem repeats (Frenkel and Trifonov, 2012) while human languages may have originated similarly, from primitive repetitive utterances of baby-babbling. In late prenatal and early postnatal ontogenesis of *H. sapiens* only very basic (ancient) muscles of speech apparatus should have appeared first, allowing for simple repetitive sounds. These, we would suggest, probably, would represent also the origin of language, first used by adult ancestral hominids, with, presumably, much simpler speech apparatus than the one possessed by modern *H. sapiens*. Common nature of evolution and languages has been understood by some researchers as well: "*If we were ever able to solve the enigma of glottogenesis, then we would possess a vital clue to the mystery of life itself."* (Danesi, 1993)

Below the ontogenetic theory of origin of language is outlined in some detail, in parallel with the recently proposed theory of origin and evolution of genomes.

Results and discussion

Origin and evolution of genomes

We start with the genome origin and evolution, described in several important details in a recent publication (Frenkel and Trifonov, 2012) The bottom line of this study is that the genomes, especially eukaryotic ones, are full of repeat sequences – tandem repetitions of simple motifs, and dispersed repetitions of longer sequences. The repeats occupy over 2/3 of

human genome (de Koning, Gu et al., 2011) Such mass of the repeats does not seem to imply a load of information. It appears rather that it is in the nature of the genomes to be constantly populated by the repeats. And the question is why it is so. The clue is given by existence of so-called triplet expansion diseases, such that certain aggressively repeating triplets spontaneously increase their copy numbers (expand), causing neurodegenerative diseases and chromosome fragility. In 1997 we have speculated that such selfishly successful sequences have been always around, including the very first steps of emerging life (Trifonov and Bettecken, 1997). Such sequences are likely to have been winners in the harsh competition in the earliest stages of life. The straightforward hypothetical scenario for the gene and genome origin would be spontaneous appearance of the repeats in the primitive early genomes, and their subsequent mutational changes towards higher sequence complexity while freshly generated repeats would continue to invade the genomes. As the very first duplex gene the aggressive sequences GCC_n and complementary GGC_n have been suggested (Trifonov, 2004a). Analysis of a very large amount of triplets of mRNA demonstrated that the aggressive triplet expansion sequences are massively present in the mRNA, in diversified but well recognizable form (Frenkel and Trifonov, 2012). In other words, the memory about original repeat expansion events still survives in the mRNA sequences. This is also true for sequences not coding for proteins (Trifonov, Volkovich et al., 2012) and for tandem repeats of all sequences.

Thus, genomes appear to emerge and evolve by spontaneous repeat expansions and concomitant diversification of the repeats to higher sequence complexity. Origin and evolution of the second script of life – language – may have had the same basic characteristics. In particular, one would expect the language in its evolution to start with simple repetitions as well, and continue into further diversification. Evolutionary intuition suggests that *"Language owes its origin to the imitation and modification,…, of various natural sounds, the voices of other animals, and man`s own instinctive cries"* (Darwin, 1871). Following this lead we now turn to the ontogenetic theory of the origin of the human language, developing it from the "instinctive cries" of babies with speech apparatus, perhaps, as simple as the one of ancestral hominids. Presumably, the "instinctive cries" have been *bona fide* communication between adult early hominids.

Ontogenesis of *H. sapiens*

In 1899 an exciting observation was made, by Ernst Haeckel (Haeckel, 2013). The consecutive stages of development of animal embryos were found to be very close to

respective evolutionary stages: “Ontogeny recapitulates phylogeny”. Later studies indicated that this is often not exactly true, but the main effect remained. For example, human embryo undergoes transformations from what at some point apparently corresponds to fish, and later vertebrae, tail, four extremities and other more advanced features sequentially appear. This, indeed, parallels fairly well the evolutionary stages of mammals: cartilaginous fish – bony fish – four-legged animals. One can imagine that some later evolutionary advances also obey in some degree the ontogeny/phylogeny trend.

In particular, the speech apparatus of the embryo and of infant is simpler at the beginning, different from the adult structures in terms of detailed anatomy and basic muscles involved. Perhaps, the same was true for adult ancestral hominids, vs. modern *H. sapiens.* For example, larynx and hyoid bone in ~4 month children drop to lower level along spine, thus, changing the spectrum of pronounced sounds. Further descending of these elements of speech apparatus leads to maturation (roughening) of voice.

One can hypothesize that the speech apparatus of, consecutively, say, *Australopithecus*, *Homo erectus*, *Homo heidelbergensis*, *Homo sapiens*, *Homo sapiens sapiens* gradually changed. “The language faculty must have evolved in the usual gradualistic way” (Pinker, 1994). E. g., L-shaped vocal tract, which is considered important for the speech versatility, has been continuously evolving, as opposed to appearing suddenly (Olson, 2003). These gradual changes may well be reflected also in the ontogeny of other elements of the speech apparatus. Other changes in the development and maturation are suggestive as well. “The toddler walks and leans with a posture which suggests primitive mankind” (Gesell, 1946) . It is very tempting to assume that the first “speaking” hominid ancestors, with only those speech muscles, which were formed at their stage of evolution, managed to pronounce primarily those consonants which were easier to pronounce. These earliest speech muscles, presumably, are also those which are formed first in the human ontogenesis. Such ontogeny-phylogeny view, that first speech of human ancestors probably was similar to first utterances of infants was suggested already by De Laguna (1927) and Danesi (1993). There is also a study of Philip Liebermann (Lieberman, 1984) who applied comparative measure on mandibles and skulls of newborn child, paleolithic Neanderthal and adult *H. sapiens*. He shows, that vocal apparatuses of child and Neanderthal share many features and suggests that vocal abilities of Neanderthals could be similar to those of modern children when they start to utter simple holophrases based on canonical babbling.

This hypothesis, on the ontogeny/phylogeny correspondence of the speech apparatus, is crucial for the view on the ontogenetic origin of language outlined in this work.

Unfortunately, the studies on the anatomy of speech in hominids and embryo are very scarce to further substantiate the view. Moreover, we do not expect that development of human individual language (or the speech) would precisely follow the way we described it here, and we only consider the general trend as we see it.

We, thus, try to fill in by relevant observations on pre-linguistic utterances and the baby babbling, and on genomic sequences, which parallel each other in many ways.

Pre-linguistic vocalization and baby-babbling

The stage before children start to speak is usually called as pre-linguistic. Entering the language is preceded with instinctive cries and sounds connected with vegetative functions, first voicing – sounds of pleasure and later vocal plays. First sounds produced by infants are nasals (such as m, n) and velars (such as g, k) later come other sounds of plosive character (stop consonants) such as p, b, d, t. (Oller 1980, Stark 1980).Essentially, altogether 4-7 month old babies can pronounce the consonants p, b, t, m, d, n, k, g, s, h, w, j (baby consonants) , while f, v, th, sh, ch, l, r are problematic at this stage of development of the speech apparatus (O`Grady, 2001). Finally, about 7th month infants start to produce repetitive sequences of successive syllables referred to as canonical babbling. The fact, that these sequences are composed from segments known from adult speech leads to impression, that the babbling is kind of primitive speech, however, this process is purely generative and unconscious (Studdert-Kennedy, 1991) created by oscillation of mandibles (MacNeilage and Davis, 1990). Nevertheless, CV patterns are not random and their repertoire depends on physical constraints of developing vocal apparatus. Importantly, the pronounceability of various consonants by babies is the same irrespective of the language environment and ethnicity of the babies (Teixeira and Davis 2002; Gildersleeve-Neumann, Davis et al., 2013; O`Grady, 2000) thus, apparently being function of the available arsenal of speech muscles, which is (largely?) the same in all babies of the same age. Some of the “difficult” sounds become pronounceable after appearance of milk teeth. Others, however, would, perhaps, depend on appearance of new speech muscles, both in phylogeny and ontogeny.

Although previous studies divided sharply pre-linguistic period and the speech period, recent studies clearly show, that there is a gradual transition between babbling and first conscious words (Vihman, Macken et al., 1985; Vihman, Ferguson et al., 1986).There are no strict points in time where the babbling ends and the words start. The two stages are overlapping. Infants pass purely generative process and gradually start to favor sounds of

ambient language. However the process is bidirectional: child adapts to surrounding language and parents co-opt new uttering patterns which sound like meaningful words.

Every mother is fascinated by the patterns uttered by the babies in their first months, initially in no connection with outer world. European mother, of course, would enthusiastically respond to spontaneous “ma-ma-ma”, thus, establishing and further consolidating the first liaison of baby words with reality. Georgian mother would react the same way to”da-da-da” (“dada” is mother in Georgian), while Swahili speaking mother (“baba”) would respond to “ba-ba-ba”. In the Table 1 the repeat words-syllables on the basis of the easily pronounceable consonants are presented, in various languages.

The repeating syllables with the baby consonants in practically all languages have “baby” meanings as illustrated in the Table 2, based on the data presented in the previous table. This “dictionary”, essentially, corresponds to major terms describing immediate environment of babies. Interestingly, cars, locomotives, bottle – rather non-ancient terms are in the list as well reflecting evolution of the environment.

The pre-linguistic “vocabulary” of babies, essentially identical for different ethnicities, largely consists of simple repeats of baby consonants alternating with vowels (pa-pa-pa, ta-ta-ta, na-na-na,…). This is very similar to the vocabulary of earliest sequence repeats in genomes (GCC-GCC-GCC, GGC-GGC-GGC,… , TG-TG-TG, A-A-A-A,… and alike). Moreover, the first primitive repeats of the baby talk soon “mutate” to their derivatives, like “mamale” (Hebrew), “nanny” etc. which also happens to nucleotide sequence tandem repeats acquiring advantageous point mutations. The reverse adaptation is also observed: often the adult words are adapted to baby`s repetitive phonetics: “co to je?” (what is this?, Czech) is pronounced as “to to ye”.

Table 1. : Sample collection of repetitive baby words in various languages. The data are collected from international language forum (see internet reference) and from our own interviews.

baba	child (Hungarian) grandmother (Czech, Polish, Russian) father (Arabic, Urdu, Mandarin, Swahili, Konkani)	nene	child (Portuguese) breast (French) deny (Czech) boy (Spanish)
bebe	pain (Lithuanian) sheep (Czech) child (Italian, French, Turkish)	nono	grandfather (Spanish)
bibi	pain (Hungarian) me (French) grandmother (Arabic) car (Russian)	nunu	food (Turkish) child (Arabic) sweet/ cute (Czech)
bobo	pain (Russian) dog (Urdu)	papa	food (Portuguese, Spanish,Italian, Filipino, Malayan) father (European,) bye-bye (Czech) mother (Japanese) grandfather (Georgian)
bubu	car (Japanese) scary (Czech) pain (English)	pipi	pee (English, Arabic,Italian bird (Czech)
dada	child (Arabic) grandfather (Polish) sleep (Czech) father (English)	popo	anus (French) feces (Spanish) buttocks (Turkish)
dede	uncle (Lithuanian) grandfather (Turkish) bottle (Filipino)	pupu	feces (Italian)
diadia	uncle (Russian)	sisi	breast (Russian)
dodo	sleep (French)	siusiu/ shushu	pee (Polish) pee (Urdu)
haha	mother (Japanese) fun (Czech)	tata	father (Polish, Czech, Russian) aunt (French) grandmother (Spanish, Portuguese)
kaka	feces (international) dirt (Arabic, Italian)	tete	bird (Arabic) father (Lithuania)
mama	mother (European, Arabic, Mandarin, Swahili, Quechua) food (Turkish, Japanese) father (Georgian)	titi	father (Japanese) breast (Russian)
meme	breast (Turkish) sleep (Filipino)	tutu	buttocks (Portuguese) penis (Filipino) locomotive(Russian) dog (Arabic)
mimi	child (Czech)	vovo	grandmother (Portuguese) grandfather (Portuguese)
mumu	scary (Filipino)	wawa	pain (Arabic) scary (Czech) water (Filipino)
nana	grandfather (Urdu) doll (Czech) sleep (Italian) grandmother (English) gother (Fijian) father (Telugu)	weewee	pee (English) penis (English)

Table 2: The most frequent meanings of the simple baby words. ("Back translations" from the Table 1.)

father	papa, baba,tata,dada, mama, nana, tete, titi
child	baba, bebe, mimi, dada, nene, nunu
grandfather	dada, dede,nana, nono, papa, vovo
grandmother	baba, bibi, nana, tata, vovo
pain	bobo,bibi, bubu, bebe,wawa
mother	mama, haha, papa, nana
scary	bubu,mumu, wawa,nunu
sleep	dada, dodo, nana, meme
pee	pipi, ,shushu, siusiu, weewee
defecation	kaka, popo, pupu
food	mama, nunu, papa
breast/milk	meme, nene, sisi
uncle	diadia, dede
aunt	papa, tata
bird	tete, pipi
dog	tutu,bobo

From the simple repeats and their mutations the evolution of language, likely, entered the next pronounceability stage – alternation of mixed consonants and vowels, like “bibika” (car, Russian) and adult words like “enemy”, or “sobaka” (dog, Russian). Similar alternations are characteristic also for protein sequences where the alternating polar and non-polar amino acid residues correspond to amphipathic alpha-helices of proteins (Zemkova, Trifonov et al. 2014).

Thus, the early language has been, likely, gradually progressing from the repeats of easiest pronounceability (alternations with the same vowel and consonant, from the list of easily pronounceable consonants) to more diverse repeats (involving difficult consonants), to heterogeneous alternations of consonants and vowels, and, finally, to a whole spectrum of words, including those which are problematic even for adults (e. g., pstruh, trout in Czech, or Mkrtchan – Armenian family name). And, again, this parallels the diversification of nucleotide and protein sequences from simple repeats to complex sequences. In this case the functional utility of the appearing new mutated forms serves as the pronounceability in language. This gradient from words easily dealt with to more complicated words, from simple to complex, is the most natural trend in evolution, common for both genetic and language scripts.

What prompts the repeats

The very first utterings of babies are not necessarily repetitive. However, the repetitions are in front lines of the audio environment of newborns. These are sucking rhythm, rocking cradle, bird songs, barking dogs, rattle toys, clock sounds, dance, music etc. The immediate repetition of whatever syllable the baby managed to pronounce is, perhaps, rewarding for the baby, who, obviously, enjoys generating repeating sounds. Various forms of alliteration are enjoyable, as we know well from poetry. Some contribution to the prompt may be given also by physiologically normal mandible oscillation, and pseudo-stuttering in child`s development.

The repeats expand in adult`s life and language. Clearly, repeats like figli-migli (flirt, Russian) or volens-nolens, villi-nilli (from Latin) are not introduced by children. The rhyme and rhythm are also forms of repetitiveness, not mentioning the refrains in the songs and chanting. Poetry and whole musical and dance realm are based on the repetitiveness (Ohno and Ohno 1986). Repeating phrases are typical for sermons and public speeches. One remarkable example is the excerpt from Martin Luter King`s speech (King, 1968):
"…if you want to say that *I was a drum major*, say that *I was a drum major* for justice. Say that *I was a drum major* for peace. *I was a drum major* for righteousness."
When the monument to M. L. King has been erected in Washington, the abbreviated, non-repetitive version of the above quote was initially used for the inscription. It caused natural protests, as the saying without the repetitions very much loses its appeal.

In biological sequences the repeats are prompted by inherent property of the tandem repeats to expand. Common mechanism of formation of the simple tandem repeats is slippage – spontaneous displacement of one strand of a repeating duplex relative to its complementary copy, and filling in the resulting single-stranded section (Wells, 1996). One reason for the repeats to stay is frequent use of tandems as copy-number dependent tuners of nearby gene activities (Trifonov, 1989; Trifonov, 2004)

Converging to linear, alphabetic scripts

The genomic script, the one we observe today, is, essentially, made from well distinguished characters (in 4-letter and 20-letter alphabets), and it, probably, appeared first in this form as well, although the alphabet of nucleic acids might have been initially simpler, involving either only strong bases (G and C), or only purines (A and G). This remains so far in the domain of mere speculations, since the "simplissimus" GCCn and GGCn (Trifonov and

Bettecken 1997; Trifonov, 2000; Trifonov, 2004) has already certain degree of organization, to which the pre-triplet life had to come by evolution from either homo-polynucleotides, or mixed sequence polynucleotides (Trifonov, 2009).

The language scripts, predominantly linear and alphabetic today, did evolve through traceable changes from various logographic systems , such as logographic-syllabic Sumerian, logographic Egyptian, and logographic Chinese, to syllabic and, finally, alphabetic modern scripts. The structural units of the evolving language scripts have been initially two-dimensional pictograms, each conveying a word, a meaning. Later it was replaced by syllables together making the words, and, finally, by letters, consonants and vowels, various combinations of which, again, make a variety of words. The latter scripts are more economical because they allow by relatively small number of symbols to express the unlimited language information. The constantly growing number of new terms, notions of expanding modernity cannot be realistically satisfied by new pictograms. Full list of Chinese characters already contains tens of thousands logograms, so that eventually it became necessary to introduce more economical writing systems, such as Japanese syllabic *hiragana* and *katakana*, of just about 50 of syllables and vowels, thus, approaching to alphabetic scripts, with 18 to 59 letters in the alphabets of various languages. The letters, minimal structural elements of the scripts, essentially, do not carry any meaning of their own, except the sound (consonant or vowel) they denote. The meaning is conveyed by combinations of the letters, very much like in nucleic acids and proteins where elementary letters correspond to simple chemical entities, nucleobases and amino acids, various combinations of which ("words", patterns) correspond to those numerous codes which are carried by biological sequences (Trifonov, 1989).

The evolution of the genetic script from "logograms" to letter sequences has been, actually, suggested as well, in the theory of compositional inheritance at the earliest stages of primitive life (Segré, Ben-Eli et al., 2001) In this theory all lipid-like amphiphilic compounds synthesized in the protocell ("metabolism first" compounds), supposedly, have been statistically shared ("bag replication") between daughter cells after division, very much like mitochondria shared today. The *composition* of the cells, thus, has been inherited, with no heteropolymer sequences at this stage. The compounds may be considered, thus, as logograms of this primitive script of heredity. Inherited in this case was a bag of the "logograms", rather than a sequence of symbols.

Language is not a privilege of *H. sapiens* only. In rudimentary forms it can be recognized in birds and animals. One remarkable example is the communication between dolphins. Their ultrasound series carry words of a sort. For example – they call each other by distinct names (King and Janik, 2013). The "names" uttered by the dolphins are all different, which suggests that these are not just instinctive simple calls, but a meaningful addressed communication. The time series of the ultrasound records appear as linear sequences of sonogram words, a script of the kind, perhaps, potentially decomposable into “alphabet” elements.

Language as biological phenomenon

We fully realize that the relation of the languages to biology is highly controversial issue. The connection, however, is very obvious, as it follows from what is described above. There are plenty of characteristics shared by the two scripts such as frequency vocabularies, rules of “pronounceability”, contrast words, Shannon N-gram extension patterns, variation of linguistic complexity and more. There are, of course, some notable differences as well. One of them is overlapping character of the multiple codes in genomes , (Trifonov, 1989) while such overlapping in languages is only rare, exotic phenomenon (like acrostics). Another important difference is strict rule of almost ideal identity in replication, while the identity in writings is, essentially, kept only for consecutive editions of canonical and authored texts. Diversity in human writings is, virtually, unlimited, while only certain degree of diversity is allowed in evolution of organisms. We believe that language and the whole corpus of writings are not only a cultural heredity but also a subject and product of biological evolution. Is there pressure of natural selection acting on ideas and canonical texts? Will the canonical texts (basic human knowledge, religious writings, and theories) eventually crystallize in frozen invariants of all cultures?

The tempting, perhaps, non-orthodox thought is: both types of texts, genetic and language scripts are products and subjects of evolution, inseparable manifestations of life, both belonging to the domain of biology, both representing heritage of *H. sapiens*, genetic and cultural. And both components of the heritage are crucially important for identity and survival of the human beings.

Acknowledgements

The work has been supported by the Czech Ministry of Education (grant MSM0021622415 and MSM0021620828) and by the project "Innovation of the General linguistics and Theory of communication in cooperation with the natural sciences" (grant CZ.1.07/2.2.00/28.0076).

References

Danesi, M. (1993): *Vico`, metaphor`, and the origin of language. Bloomington:* Indiana Univeristy Press.

Darwin, Ch. (1871): *The descent of man, and selection in relation to sex.* London, Murray.

de Koning, A. J. - Gu, W. - Castoe, T. A. - Batzer, M. A. - Pollock, D. D. (2011): Repetitive elements may comprise over two-thirds of the human genome. *PLoS genetics,* **7**(12): e1002384.

De Laguna,G. (1927): *Speech: its function and development*. Oxford, Yale Univ. Press.

Frenkel, Z. M. - Trifonov, E. N. (2012): Origin and evolution of genes and genomes. Crucial role of triplet expansions. *J Biomol Struct Dyn,* 30(2), p.201-210.

Gesell, A. (1946). *The ontogenesis of infant behavior.* In: Carmichael, Leonard (eds.), Manual of child psychology. John Wiley & Sons Inc., p. 295-331.

Gildersleeve-Neumann, Ch.,E.- Davis, B.L. – Macneilage, P. F. (2013): Syllabic patterns in the early vocalizations of Quichua children, *Applied Psycholinguistics,* 34/01, p.111-134.

Gould, S. J. (1977): *Ontogeny and Phylogeny,* Belknap Press of Harvard University Press.

Haeckel, E. H. P. (2013): *The Riddle of the Universe at the Close of the Nineteenth Century.* General Books LLC. Is this original reference, date?

King, M.L.(1968): sermon in Atlanta

King, S. L.- Janik, V. M. (2013): Bottlenose dolphins can use learned vocal labels to address each other. *Proceedings of the National Academy of Sciences* **110**(32), p. 13216-13221.

Lieberman, P. (1984): *The biology and evolution of language.* Harvard University Press

MacNeilage, P. F. - Davis B. (1990): Acquisition of speech production: Frames, then content.In: Jeannerod, Marc (eds.) *Attention and performance 13: Motor representation and control.* England: Lawrence Erlbaum Associates, Inc, p. 453-476.

O`Grady, W. (eds.) (2001): *Contemporary linguistics: an introduction*. St Martins Press (NY).

Ohno, S. - Ohno, M. (1986): The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition. *Immunogenetics,* 24(2), p.71-78.

Oller, D. K. (1980): The emergence of the sounds of speech in infancy. *Child phonology* **,** 1, p. 93-112.

Olson, S. (2003): *Mapping Human History: Unravelling the Mystery of Adam and Eve.* Bloomsbury.

Pinker, S. (1994): *The language instinct: The new science of language and mind,* Penguin UK.

Segré, D. - Ben-Eli, D.- Deamer, D. W. - Lancet, D. (2001): The lipid world. *Origins of Life and Evolution of the Biosphere*, 31(1-2), p. 119-145.

Stark, R. E. (1980): Stages of speech development in the first year of life. *Child phonology*, **1**, p. 73-90.

Studdert-Kennedy, M. (1991): Language development from an evolutionary perspective. *Biological and behavioral determinants of language development,* p. 5-28.

Teixeira, E.R. - Davis, B.L. (2002): Early Sound Patterns in the Speech of Two Brazilian Portuguese Speakers, *Language and Speech,* 45/2, p. 179-204.

Trifonov, E.N. - Bettecken, T. (1997): Sequence fossils, triplet expansion, and reconstruction of earliest codons. *Gene,* 205(1), p. 1-6.

Trifonov, E. N. (1989): The multiple codes of nucleotide sequences. *Bull Math Biol,* 51(4),p. 417-432.

Trifonov, E. N. (2000): Consensus temporal order of amino acids and evolution of the triplet code. *Gene,* 261(1), p.139-151.

Trifonov, E. N. (2004a): The triplet code from first principles. *J Biomol Struct Dyn,* 22(1), p. 1-11.

Trifonov, E. N. (2004b): Tuning function of tandemly repeating sequences: a molecular device for fast adaptation. *Evolutionary theory and processes: Modern horizons,* Springer**:** p. 115-138.

Trifonov, E. N. (2009): The origin of the genetic code and of the earliest oligopeptides. *Research in microbiology***,** 160(7), p. 481-486.

Trifonov, E. N. - Volkovich, Z. - Frenkel. Z. M. (2012): Multiple levels of meaning in DNA sequences, and one more. *Annals of the New York Academy of Sciences,* 1267(1), p. 35-38.

Vihman, M. M.- Ferguson, Ch. A.- Elbert,M. (1986) : Phonological development from babbling to speech: Common tendencies and individual differences. *Applied Psycholinguistics*, 7/01, p.3-40.
Vihman, M.M. - Macken,M.A. - Miller,R. – Simmons, H. - Miller, J.(1985) : From Babbling to Speech: A Re-Assessment of the Continuity Issue. *Language*, 61(2) , p. 397-445.

Wells, R. D. (1996): Molecular basis of genetic instability of triplet repeats. *Journal of Biological Chemistry***,** 271(6), p. 2875-2878.

Zemkova, M. - Trifonov, E. N. - Zahradnik, D. (2014): One common structural feature of "words" in protein sequences and human texts. *J Biomol Struct Dyn*, 32(7), p. 1085-1091.

Internet references
Language forum- Baby-talk vocabulary:
http://forum.wordreference.com/showthread.php?t=1060391 (Accessed 20 September 2014)

Ostatní publikace

Zemková M. (2008) „Nová divočina“ v Praze – její biodiverzita a estetika. In K. Stibral, B. Binka a O. Dadejík (Eds.), Krása, krajina, příroda II. Brno: Muni Press.

Zemková M. (2011) New wilderness.In Petr Gibas, Karolína Pauknerová and Marco Stella et al. Non-humans in Social Science: Animals, Spaces, Things. Cerveny Kostelec: Pavel Mervart. ISBN 978-80-7465-010-9.