



**Biofizikální ústav**  
Akademie věd České republiky  
**doc. RNDr. Eduard Kejnovský, CSc.**  
Oddělení vývojové genetiky rostlin  
Královopolská 135, 612 00 Brno  
tel. 541 517 203, fax: 541 240 500  
e-mail: kejnovsk@ibp.cz  
[www.ibp.cz/labs/DPDG](http://www.ibp.cz/labs/DPDG)

---

Oponentský posudek dizertační práce doktorandky Ing. **Michaely Nohejlové Zemkové**  
**“Lingvistické přístupy v genomice a lingvistická metafora v biologii“**

Dizertační práce Ing. Michaely Nohejlové Zemkové se zabývá lingvistickými přístupy v genomice, tedy snaží se o pohled na genomy a jejich části jako na text psaný nějakým jazykem. Práce se dosti vymyká pracem, které jsem dosud posuzoval, chvílemi jsem tak měl pocit, že jde o bláznivý nápad či úlet, chvílemi pak, že se jedná o vtipné nápady. Skutečnost, že výsledky byly publikovány pouze v jednom článku v impaktovaném časopise, podporovala první pocit, avšak protože znám originalitu a provokativnost myšlenek profesora Trifonova i skvělý smysl pro humor profesora Flegra, přikláním se spíše ke druhému vysvětlení. Práce sestává ze čtyř částí, kdy každá z nich obsahuje úvod, metodiku, výsledky, diskusi a závěr, vše v rozsahu 42 stran. Těmto čtyřem částem odpovídají rukopisy čtyř publikací autorky, které následují, přičemž jedna byla publikována v impaktovaném mezinárodním biologickém časopise, dvě v českém lingvistickém časopise a jedna odeslána.

Než přikročím k uvedeným čtyřem částem podrobněji, rád bych **obecně** řekl, že zdaleka ne celá práce je napsána čtivě, některé pasáže, zejména týkající se metodiky, jsou těžko srozumitelné, text obsahuje mnoho překlepů (jen jako příklad: 2x na str. 42 „repetitivně“ místo repetitivní“, str. 42 řádek 7 chybí mezera mezi slovy, str. 42 – jednou Alu sekvence s velkým „A“, jindy s malým „a“). Vadilo mi i špatné používání teček na konci věty před závorkou místo za ní, dále dvě citace lze dát za sebe do jedné závorky, není potřeba, aby každá měla svoji závorku, neznámé pro mne bylo slovo „vylišit“ ve spojení „lze vylišit

proteomy“ (opakovaně na str. 38). Některé části si dokonce snad i protiřečí – konkrétně v druhé části porovnávající genomový text s textem lidského jazyka se nejprve uvede, že obojí je umělá konstrukce/konvence a že analogie tedy nelze hledat, a přitom vzápětí je prezentována analýza porovnávající oba typy textů. Je rovněž škoda, že text obsahuje pouze dva obrázky, více obrázků by práci oživilo a zajisté jí prospělo.

Nyní bych postupně okomentoval všechny čtyři části práce a současně položil některé otázky.

### **Část první (redukce peptidových slov u parazitů):**

1. v Závěru na str. 29 se uvádí, že existují čtyři typy důkazů pro hypotézu redukce bohatosti peptidových slov u parazitů a jako jeden z nich autorka uvádí neredukovaný slovník u prvoka *Perkinsus marinus* (parazitujícího na ústřicích) – jak to dokazuje uvedenou hypotézu? Je to myšleno tak, že lze tuto výjimku, tedy parazita s neredukovaným slovníkem, vysvětlit nepřítomností MHC u hostitele? A tedy nepotřebností redukovat slovník, aby unikli obraně?
2. Text popisující lingvistickou analýzu genetických sekvencí (str. 17-22) a související obrázek 1 jsou těžko stravitelné. Nepochopil jsem například, čím je délka slova 5 písmen tak výjimečná.
3. Jak si vysvětlujete rozdíly v diverzitě tetra-, penta- a hexapeptidů u parazitů (str. 28)? Proč je diverzita některých zvýšená a jiných snižená?

### **Část druhá (slova v proteinech a lidském textu):**

1. Souhlasím s tvrzením, že jak polarita aminokyselin, tak hlásky (samohlásky a souhlásky) jsou výsledkem konsenzu a nejsou přirozené a tudíž nemá smysl mezi nimi hledat souvislost. Proč je tedy toto trefné tvrzení následováno další analýzou? Její výsledek navíc není nějak atraktivní, neboť nebyl prokázán rozdíl mezi parazity a volně žijícími organizmy ve využití amfipatických slov.
2. Na str. 37 je uvedeno, že odstraněním mezer mezi slovy vznikají neexistující slova – jak to ovlivní výsledky analýz a případně jejich závěry?
3. Potěšil mě příklad posloupnosti nejčastějších slov v proteomu jednoho zkoumaného organismu (str. 38), neboť jeden příklad je lepší než stránky popisujícího textu.

4. V případě obrázku 2 (str. 39) je pro čtenáře těžko pochopitelné, co je na osách.

5. Na str. 39 je uvedeno, že v proteinech se střídají polární a nepolární (amfipatické) řetězce zatímco v řeči se střídají souhlásky a samohlásky. V řeči je důvodem snadná vyslovitelnost, důvody pro střídání v proteinech jsou však jiné – jaký selekční tlak by mohl způsobit takové vlastnosti?

### **Část třetí (slova v lidském genomu):**

1. Na str. 43 je uvedeno, že vzorek lidské DNA sekvence byl dlouhý pouze 10Mb (psáno 10MB jako kdyby šlo o MegaByty). Ptám se – není to málo? Zejména vzhledem k heterogenitě lidské sekvence, existenci izochor či dlouhých repetit. Analýza je triviální až naivní a zkušenému bioinformatikovi by neměla zabrat moc času.

2. Na str. 44 je uvedeno, že „velký počet jednoduchých tandemových repetit v genomu naznačuje, že se jedná o generátory“ – divný závěr, z čeho plyne?

### **Část čtvrtá (genom a jazyk):**

1. Na str. 46 řádek 7 se uvádí, že „původní ... repetitivní povaha genomu je stále patrná a rekonstruovatelná z mRNA“. Jak rekonstruovatelná a co je vůbec repetitivní povahou myšleno? Proč číst repetitivní povahu z mRNA a nikoliv z DNA? Repetice se přece nacházejí více v intronech než exonech a introny v mRNA již nejsou obsaženy.

2. Na str. 48 se lze dočíst, že výsledkem je „předložený slovník dětských repetitivních slov“ – opět primitivní výsledek, v metodice se uvádí, že data byla čerpána z webu a osobních interview autorů.

3. Na str. 49 v Závěru najdeme větu, že „Otázka po vzniku života je tudíž podobně zastřené povahy jako otázka po vzniku jazyku.“ Věta zde působí divně, jako naivní závěr, jde přece o dvě věci zkoumané zcela odlišnými přístupy.

### **Dovolil bych si položit doktorandce pár obecnějších otázek jako základ pro širší diskusi, která k důstojné obhajobě zajisté patří:**

1. Genomy jakých organismů by byly dle vás nejvhodnější pro hledání prvních nejprimitivnějších slov v nukleových kyselinách? Byly by to genomy pokročilých organismů

jako je člověk (jak to děláte vy) anebo jednoduché genomy molekulárních fosilií jako je třeba fág Qbeta?

2. Domnívám se, že nejen redukovaný počet slov ve slovníku genomu parazitů, ale také nižší počet proteinů může odrážet parazitický způsob života. Pokud vím, tak paraziti mají často malé genomy, méně repetic, méně duplikací a tudíž bych čekal i menší diverzitu genů a snad i proteinů? Byla tato skutečnost ve vašich analýzách nějak zohledněna?

3. Jak často ve vědecké literatuře se používají lingvistické metody při studiu genomů?

**Závěrem musím konstatovat**, že i navzdory četné kritice celkově hodnotím dizertační práci Ing. Michaely Nohejlové Zemkové jako zdařilou. Doktorandka získala značné množství zajímavých (a publikovaných) výsledků, což svědčí nejen o bioinformatických dovednostech, ale i o její schopnosti získané výsledky správně interpretovat a při tom zúročit i svoje teoretické znalosti. Protože studentka prokázala schopnost tvůrčí vědecké práce a práce splňuje požadavky kladené na dizertační práci v daném oboru, jednoznačně ji doporučuji k obhajobě.

V Brně, dne 15. srpna 2016

Eduard Kejnovský