FUZZY CLASSIFICATION FOR ANALYSIS OF SPECIES MEMBERSHIP IN DNA BARCODING

Jiří Těthal

Master Degree Programme (2), FEEC BUT

E-mail: xtetha00@stud.feec.vutbr.cz

Supervised by: Helena Škutková

E-mail: skutkova@feec.vutbr.cz

Abstract: DNA Barcoding is an approach that seeks to identify species using short sequences, mostly mitochondrial DNA. Biological systems, by their nature are stochastic, boundaries between sequences are fuzzy, hence the need to use alternative methods that can be used for automated evaluation. Fuzzy set theory and fuzzy logic provides us with another way of view of the modeling uncertainty. The combination of fuzzy logic with DNA barcoding provides a more natural classification of biological sequences using fuzzy boundaries between species. This article demonstrates the use of fuzzy classification of DNA barcoding in experimental sample sequences.

Keywords: DNA Barcoding, Fuzzy, Membership function

1. ÚVOD

DNA Barcoding je technika, umožňující identifikaci a klasifikaci neznámých vzorků pomocí krátkého úseku sekvence, obvykle mitochondriální DNA. Z hlediska evoluce dochází v mitochondriální DNA k mnohem více mutacím než v jaderné DNA a lze díky ní odlišit i velmi blízké druhy na poměrně malém úseku (cca 650 bp). Molekulární evoluce je ale ze své podstaty stochastický proces, tj. hranice mezi sekvencemi jsou neostré. Proto je k automatizovanému vyhodnocení výhodné využít nedeterministických metod, jako jsou pravděpodobnostní modely nebo fuzzy teorie množin a fuzzy logika. Spousta fyziologických a evolučně významných procesů v organismu je náhodných nebo nejistých, pravděpodobnostní popis všech jejich příčin je vždy neúplný a stává se tak pouze odhadem. Kombinace fuzzy logiky společně s DNA barcodingem je biologicky přirozenější klasifikací sekvencí využívající neostrých hranic mezi jednotlivými druhy. Tento článek demonstruje využití fuzzy klasifikace v DNA barcodingu na experimentálním vzorku sekvencí ryb, konkrétně *Actinopterygii of Churchill*, získaných z veřejně přístupné databáze barcode sekvencí BOLD (http://www.barcodinglife.com). Set obsahuje 180 sekvencí 17 druhů.

2. FUZZY LOGIKA

Pojem fuzzy logika poprvé použil profesor Lotfi A. Zadeh z Kalifornské univerzity v Berkeley roku 1965 v článku Fuzzy sets. Inf. & Control, zabývajícím se rozvojem modifikované teorie množin [3]. Slovo fuzzy znamená neostrý, matný, mlhavý, neurčitý, vágní. Odpovídá tomu i to, čím se fuzzy teorie zabývá: snaží se pokrýt realitu v její nepřesnosti a neurčitosti. Je nástrojem pro matematický popis vágních a nepřesných pojmů. V klasické teorii množin prvek do množiny patří (úplné členství v množině) nebo nepatří (žádné členství v množině). Fuzzy množina je množina, která kromě úplného nebo žádného členství připouští i členství částečné. To znamená, že prvek patří do množiny s jistou mírou členství, která je vyjádřena stupněm příslušnosti, vyjadřujícím příslušnost k množinám v rozmezí od 0 do 1, včetně obou hraničních hodnot. Fuzzy logika tak umožňuje matematicky vyjádřit pojmy jako „málo“ nebo „hodně“. Přesněji, umožňuje vyjádřit částečnou příslušnost k množině.

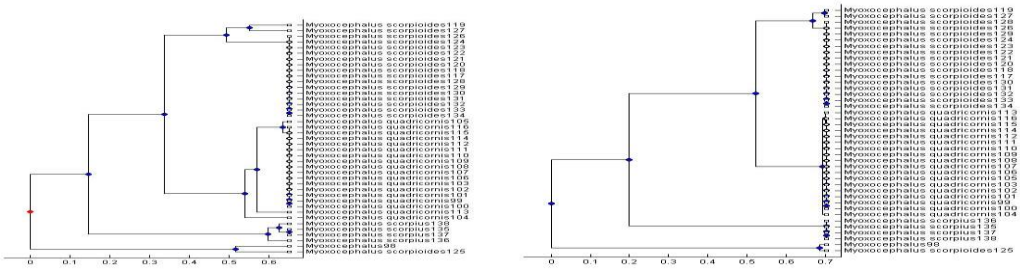
3. KLASIFIKACE ORGANISMŮ

Evoluční vzdálenost u ryb *Actinopterygii* je počítána dvěma způsoby. Prvním je euklidovská vzdálenost z vypočítaných denzit DNA (s délkou okna 30)[2] a druhým je výpočet pomocí běžně používané metody Jukes-Cantor (JC)[3]. Tyto vzdálenosti jsou pak normalizovány a jdou na vstup fuzzy funkce příslušnosti, kterou jsou zkorigovány. Výsledkem je hodnota FMF (fuzzy funkce příslušnosti). Tato funkce je definované v detailu podle následující rovnice 1 [4]:

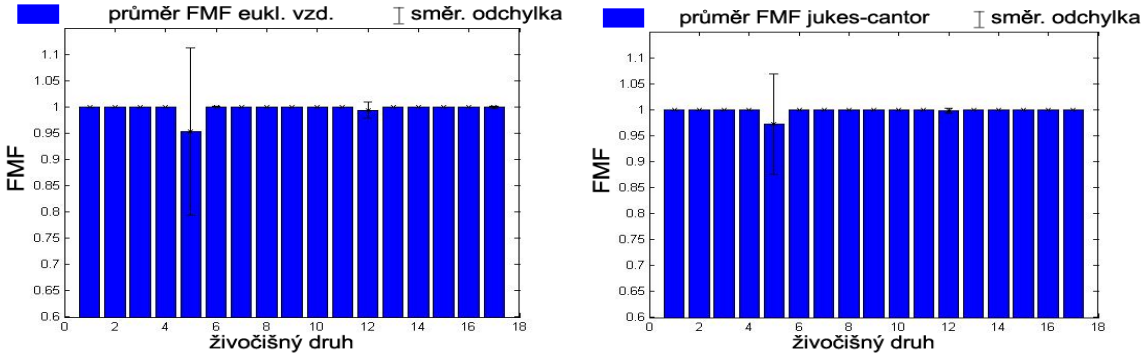
$$f(x;\theta) = \begin{cases} 1, & x < \theta_1 \\ 1 - 2\left(\frac{x-\theta_1}{\theta_2-\theta_1}\right)^2 & \theta_1 \le x \le \frac{\theta_1+\theta_2}{2} \\ 2\left(\frac{x-\theta_2}{\theta_2-\theta_1}\right)^2 & \frac{\theta_1+\theta_2}{2} \le x \le \theta_2 \\ 0, & x > \theta_2 \end{cases} \quad (1)$$

kde x je definováno jako vzdálenost dvou sekvencí nebo vzdálenost sekvence od referenčních sekvencí jednotlivých druhů. Dva parametry Φ_1 a Φ_2 je třeba odhadnout dle konkrétního souboru dat. Jedná se o maximální vnitrodruhovou a minimální mezidruhovou vzdálenost sekvencí. V našem případě byl na základě častého používání u DNA barcodingu vybrán 5. percentil pro vzdálenost v rámci druhu a 95. pro mezidruhovou vzdálenost.

Na Obrázku 1 vidíme sestrojené dva dendrogramy s výpočtem vzdáleností sekvencí pomocí JC. První dendrogram je bez použití FMF. Jedná se zde o sekvence 3 druhů patřících k jednomu rodu, což je z prvního dendrogramu nezřetelné. Na druhém je ale vidět, že došlo ke značnému zjednodušení. Vzdálenosti mezi jednotlivými zástupci stejného druhu se zkrátily, či úplně vymazaly, naopak mezi druhy jsou vzdálenosti větší. Zde je zřetelné rozdělení jednotlivých druhů díky použití FMF. Mimo tyto 3 hlavní skupiny leží ještě dvě sekvence, které byly vzdálenější oproti ostatním pro svou lišící se délku.



Obrázek 1: Srovnání dendrogramů – vlevo metodou JC bez použití FMF, vpravo s jejím použitím



Obrázek 2: Průměrné hodnoty FMF pro jednotlivé druhy s vyznačenou směrodatnou odchylkou

Dále jsou porovnány výsledky FMF pro jednotlivé druhy, na Obrázku 2 vlevo pro euklidovskou vzdálenost a vpravo pro vzdálenost pomocí JC. Jsou zde vyneseny průměrné hodnoty FMF v rámci jednotlivých druhů s vyznačenou směrodatnou odchylkou. Tyto hodnoty jsou přehledněji zaznamenány i v Tabulce 1. Dále je zde počet sekvencí jednotlivých druhů a počty sekvencí, které

v rámci druhu neodpovídají úplné příslušnosti (FMF≠1). Příslušnost ke svému druhu menší než 90 % mají u obou metod pouze 2 sekvence. Avšak při sestrojení dendrogramu byly i tyto 2 sekvence zařazeny ke svému správnému druhu.

Tabulka 1: Porovnání JC a Euklidovké vzdálenosti

	Jukes-Cantor					Euklidovská vzdálenost				
název druhu	průměr FMF	směr. odch.	počet sek-vencí	FMF<0,9	FMF≠1	průměr FMF	směr. odch	počet sekvencí	FMF<0,9	FMF≠1
Ammodytes hexapterus	1	0	3	0	0	1	0	3	0	0
Catostomus commersonii	1	0	8	0	0	1	0	8	0	0
Cottus cognatus	1	0	7	0	0	1	0	7	0	0
Cottus ricei	1	0	1	0	0	1	0	1	0	0
Culaea inconstans	0,972	0,096	48	2	6	0,958	0,139	48	2	7
Cyclopterus lumpus	1	0	17	0	1	1	0	17	0	1
Gymnocanthus tricuspis	1	0	11	0	0	1	0	11	0	0
Lota lota	1	0	1	0	0	1	0	1	0	0
Lumpenus fabricii	1	0	1	0	0	1	0	1	0	0
Myoxocephalus	1	0	1	0	0	1	0	1	0	0
Myoxocephalus quadri-	1	0	18	0	0	1	0	18	0	0
Myoxocephalus scorpio-	0,998	0,005	18	0	2	0,995	0,013	18	0	2
Myoxocephalus scorpius	1	0	4	0	0	1	0	4	0	0
Pungitius pungitius	1	0	21	0	0	1	0	21	0	1
Rhinichthys cataractae	1	0	14	0	0	1	0	14	0	2
Stichaeus punctatus	1	0	3	0	0	1	0	3	0	0
Thymallus arcticus	1	0	4	0	0	0,998	0,002	4	0	1

4. ZÁVĚR

Využití fuzzy funkce příslušnosti společně s barcodingem se ukázalo být užitečné především u klasifikace velmi podobných druhů, kdy zmenšuje rozdíly mezi jednotlivými zástupci v rámci druhu a tím zvětšuje rozdíly mezi jednotlivými druhy. Z analýzy dále vyplývá, že použití metody JC pro výpočet vzdáleností mezi sekvencemi dává mírně lepší výsledky než výpočet euklidovské vzdálenosti z denzit DNA (vždy však záleží na volbě parametrů u samotného výpočtu denzit a na typu sekvencí). Rozdíl před a po použití FMF ve vykreslených dendrogramech je patrný ve zjednodušení struktury a jednodušší interpretaci. Všechny sekvence byly přiřazeny správně ke svému druhu, pouze u dvou byla hodnota podobnosti nižší a to z důvodu kratších sekvencí o 1/6.

REFERENCE

[1] ZHANG, A.-B., C. MUSTER, H.-B. LIANG, C.-D. ZHU, R. CROZIER, P. WAN, J. FENG a R. D. WARD. A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Molecular Ecology* [online]. 2012, roč. 21, č. 8, s. 1848-1863 ISSN 09621083. DOI: 10.1111/j.1365-294X.2011.05235.x.

[2] MADĚRÁNKOVÁ, D.; PROVAZNÍK, I. Similarity/ Dissimilarity Analysis of COI Mitochondrial Gene of Chosen Bird Species Based on Nucleotide Density. In *10th International Conference on Information Technology and Application in Biomedicine*. Korfu: IEEE, 2010. s. 1-4. ISBN: 978-1-4244-6560-6.

[3] JUKES, TH; CANTOR, CR (1969) Evolution of protein molecules. In Munro HN, editor, *Mammalian Protein Metabolism*, pp. 21-132, Academic Press, New York.

[4] ZADEH, L. A.: Fuzzy sets. Inf. & Control, 8, 1965, s. 338-353

[5] YUAN J, SHI HB, LIU C (2008) Construction of fuzzy membership functions based on least squares fitting. Control & Decision, 23, 1263–1271.