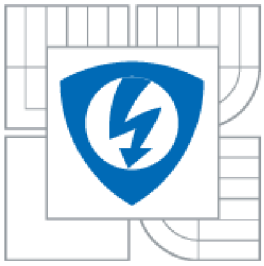
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

FUZZY KLASIFIKACE DNA SEKVENCÍ

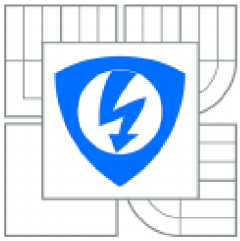
FUZZY CLASSIFICATION OF DNA SEQUENCES

DIPLOMOVÁ PRÁCE
MAGISTER'S THESIS

AUTOR PRÁCE — Bc. Jiří Těthal
AUTHOR

VEDOUCÍ PRÁCE — Ing. Helena Škutková
SUPERVISOR

BRNO, 2013



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor
Biomedicínské inženýrství a bioinformatika

Student:	Bc. Jiří Těthal	***ID:***	119750
Ročník:	2	***Akademický rok:***	2012/2013

NÁZEV TÉMATU:

Fuzzy klasifikace DNA sekvencí

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši o možnostech využití fuzzy logiky v klasifikaci biologických sekvenčních dat. Zaměřte se zejména na problematiku taxonomického třídění sekvencí a popisu společných DNA motivů. 2) Formulujte různá klasifikační kritéria a vhodné fuzzy operátory na základě známých klasifikovaných DNA sekvencí. Na základě těchto kritérií navrhněte klasifikační algoritmus pro posouzení podobnosti a příbuznosti DNA sekvencí. 3) Vytvořte testovací databázi DNA sekvencí z veřejných databází pro možnosti aplikace klasifikačních algoritmů. 4) Otestujte vlastní navržené klasifikační algoritmy využívající fuzzy logiku v programovém prostředí Matlab s Bioinformatickým toolboxem. 5) Statisticky vyhodnoťte a diskutujte úspěšnost klasifikace na základě různých kritérií a v závislosti na různých typech DNA sekvencí.

DOPORUČENÁ LITERATURA:

[1] NASSER, Sara, Adrienne BRELAND, Frederick C. HARRIS a Monica NICOLESCU. A fuzzy classifier to taxonomically group DNA fragments within a metagenome. In: Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society, 2008.: NAFIPS 2008. New York, 2008, s. 1-6.
[2] GARCIA, Fernando, Francisco J. LOPEZ, Carlos CANO a Armando BLANCO. Study of fuzzy resemblance measures for DNA motifs. IEEE International Conference on Fuzzy Systems. Korea 2009, s. 1175-1180.

Termín zadání:	11.2.2013	***Termín odevzdání:***	24.5.2013

Vedoucí práce: Ing. Helena Škutková
Konzultanti diplomové práce:

prof. Ing. Ivo Provazník, Ph.D.
Předseda oborové rady

ABSTRAKT

Práce se zabývá Fuzzy klasifikací sekvencí DNA.
V první části práce jsou teoreticky shrnuty informace o Fuzzy logice a metodách jejího využití v klasifikaci biologických sekvenčních dat.
Druhá část se již prakticky zabývá řešením klasifikačního algoritmu pro posouzení podobnosti sekvencí. Konkrétně pro rozdělení kódujících a nekódujících částí sekvence a využití fuzzy klasifikace v DNA barcodingu.

ABSTRACT

The work deals with the fuzzy classification of DNA sequences.
In the first part the theory summarized information about Fuzzy logic and methods of its use in the classification of biological sequence data.
The second part is practically deal with the classification algorithm for assessing the similarity of sequences. Specifically, the dividing of coding and non-coding parts of the sequence and the use of fuzzy classification in DNA barcoding.

Klíčová slova

Fuzzy, DNA, DNA barcoding, BOLD, beta-globin, cytochrom c-oxydáza, sekvenční motiv, genom, EMBL, GC obsah, exon, intron.

Keywords

Fuzzy, DNA, DNA barcoding, BOLD, beta-globin, cytochrome c-oxidase, sequence motif, genome, EMBL, GC content, exon, intron.

Bibliografická citace

TĚTHAL, J. Fuzzy klasifikace DNA sekvencí. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2013. 78 s. Vedoucí diplomové práce Ing. Helena Škutková.

Prohlášení

Prohlašuji, že svoji diplomovou práci na téma Fuzzy klasifikace DNA sekvencí jsem vypracoval samostatně pod vedením vedoucí diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících, autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne 23. května 2013 .....

podpis autora

Poděkování

Děkuji vedoucí diplomové práce Ing. Heleně Škutkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

V Brně dne 23. května 2013 .....

podpis autora

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Účinná klasifikace biologických sekvencí je nezbytná k lepšímu pochopení jejich funkce. Uplatňuje se ve třídění organismů, ale slouží i pro jejich identifikaci, nebo určení evolučního vývoje druhů. Při získání nové sekvence je vhodné stanovit její vzájemnou podobnost k jiným známým sekvencím, protože se z ní dá usuzovat její funkce, struktura i taxonomické zařazení. Biologické systémy jsou ale ze své podstaty stochastické a obsahují řadu nejistých nebo těžce definovatelných procesů. Proto i každý závěr stanovený na základě podobnosti primární struktury biologických sekvencí je pouze odhad, který je tak přesný, jak úplný máme popis všech procesů vedoucích k jejich vzniku. Mezi časté nedostatky však patří právě neúplnost údajů při získávání vzorků, dále nedostatek expresivity některých znaků, nejasné hranice mezi taxonomickými třídami a v neposlední řadě hlavně neznalost všech mutagenních vlivů, které mohly na sekvence působit. V těchto situacích je potřeba využít alternativních metod, které mohou být použity k automatizovanému vyhodnocení.

Téměř všechny problémy jsou v bioinformatice řešeny deterministicky. Jsou definované pevně dané funkce, které jsou řešené pomocí optimalizace. Mnoho dynamických procesů, jako je například regulace genové exprese, je také řešeno deterministicky pomocí diferenciálních rovnic.

Fuzzy teorie množin a fuzzy logika nám poskytuje jiný způsob pohledu na modelování nejistoty a nabízí nám široké spektrum výpočetních nástrojů k usnadnění rozhodování. Patří mezi přirozené způsoby, jak modelovat nejednoznačné události, které si potřebujeme zdůvodnit. Ukazuje se proto jako účinný nástroj pro modelování a analýzu biologických dat či systémů. V mnoha výzkumech byla tato účinnost již prokázána při řešení širokého spektra biologických problémů nalezených v bioinformatice, biomedicínské inženýrství a výpočetní biologii. Bylo prokázáno, že mnohé náhodné nebo nejisté procesy v organismu jsou fyziologické a evolučně významné pro rozvoj a funkce živých organismů. [32]

Tato práce se proto snaží přiblížit pojem fuzzy logiky a také metod, které ji využívají při analýze biologických dat. Konkrétně se zabývá fuzzy klasifikací DNA sekvencí a to jak klasifikací funkce DNA, tedy identifikací kódujících oblastí, tak klasifikací taxonomickou. Je zde navržena a představena řada klasifikačních kritérií pro rozlišení kódujících a nekódujících sekvencí, jako jsou obsah GC, frekvence jednotlivých oligonukleotidů či DNA motivy. Dále je zde pojednáno o možnostech využití fuzzy klasifikace v DNA barcodingu pro zpřesnění identifikace druhů. Tyto postupy jsou pak prakticky použity a ověřeny při analýze experimentálních vzorků sekvencí získaných z veřejně přístupných databází.

1. Fuzzy

1.1. Fuzzy logika

Fuzzy logika (česky též mlhavá logika (fuzziness = mlhavost)) je podobor matematické logiky, který je odvozený od teorie fuzzy množin, v němž se logické výroky ohodnocují mírou pravdivosti.
Liší se tak od klasické výrokové logiky, která používá pouze dvě logické hodnoty – pravdu (označována 1) a nepravdu (označována 0). Fuzzy logika může operovat se všemi hodnotami z intervalu <0; 1>, kterých je nekonečně mnoho. Náleží mezi vícehodnotové logiky. Fuzzy logika pracuje s vágními pojmy, které mají neostré hranice. Setkáváme se s nimi v běžném životě. Jedná se například o pojmy: menší, větší, více menší, vlažný, téměř studený apod. Otázkou zůstává, co ještě patří do popsané množiny a co již ne. S tímto problémem se ale můžeme setkat již v tzv. Paradoxu z antického Řecka: Mějme malou hromadu kamení. Pokud přidáme jeden kámen, dostaneme opět malou hromadu. Tedy každá hromada kamení je malá. Jak je vidět, problém zůstává s hraničními body a použití klasických množin tedy nepřichází v úvahu. Jedním z řešení jsou právě fuzzy množiny. Fuzzy logika může být pro řadu reálných rozhodovacích úloh vhodnější než klasická logika, protože usnadňuje návrh složitých řídících systémů. Pojem fuzzy logika se poprvé objevil roku 1965 v článku [34], jehož autorem byl profesor Lotfi A. Zadeh z Kalifornské univerzity v Berkeley. Vznikla rozvojem modifikované teorie fuzzy množin. 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 jakýmsi nástrojem pro matematický popis vágních a nepřesných pojmů. Striktní popis vede k popisu skutečnosti pouze pomocí dvouprvkové množiny {0,1}. Pokud problém nelze jednoznačně určit, rozkládá se na menší podproblémy, ale poté lze použít opět jen dvouprvkovou množinu. V případech, kdy je již nemožné nebo neúnosné takto problém rozdělit, dopouštíme se jisté chyby a tím je dán odklon od reality. V roce 1966 L. A. Zadeh ve svém článku o nových směrech analýzy komplexních systémů formuloval tzv. princip inkompatibility: "Roste-li složitost systému, klesá naše schopnost formulovat přesné a významné soudy o jeho chování, až je dosaženo hranice, za níž jsou přesnost a relevantnost prakticky vzájemně se vylučující charakteristiky." Využívá se například v umělé inteligenci, v matematice, v logické analýze jazyka i v průmyslu (fuzzy regulátory) a v kvantové fyzice. [11], [15], [18], [22], [34]

1.1.1. Stupeň příslušnosti

V klasické teorii množin prvek do množiny buďto 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. Funkce příslušnosti pak přiřazuje ve fuzzy logice 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. V případě, že prvky universa jsou reálná čísla, existuje více možností matematického popisu průběhu růstu respektive klesání hodnot stupně příslušnosti. Pro prvky universa v okolí hraničních bodů by mělo platit, že čím víc se blíží prvky universa k hraničním bodům, tím pomaleji roste (klesá) hodnota stupně příslušnosti. [18], [20]

Stupeň příslušnosti je často zaměňován s pravděpodobností. Tyto pojmy jsou ale rozdílné. Fuzzy hodnota je přiřazena funkcí příslušnosti k vágně definovaným množinám a nepředstavuje pravděpodobnost nějakého jevu. Tzn. pokud prvek patří do množiny s 80% pravděpodobností, pak se v ní vyskytne v průměrně 80 případech ze 100, ale v každém jednom případě v množině buď bude, nebo ne. Naopak, pokud má prvek příslušnost k množině 80 %, znamená to, že z osmi desetin do množiny patří. Takže do množiny patří vždy, ale jen z části. [18]

Jinou disciplínou vědy, kde se zdá, že se využívá principů fuzzy logiky, je kvantová fyzika, která také počítá s tím, že mohou existovat i stavy, u kterých je výsledek měření předpověditelný pouze v rámci pravděpodobnosti. [20]

Příkladem mohou být 4 l vody v 10 litrové nádobě. Máme dvě fuzzy množiny: Plná nádoba a Prázdná nádoba. Takto částečně naplněná nádoba pak přísluší z 0,6 k Prázdné nádobě a z 0,4 k Plné nádobě. [18]

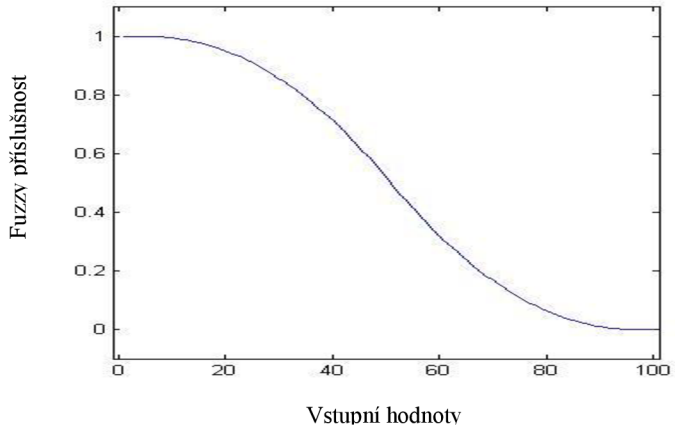
1.1.2. Druhy fuzzy funkcí

Na následujících obrázcích jsou zobrazeny různé případy funkcí, které se využívají k vyhodnocování ve fuzzy logice. [8]

Z funkce

Na Obr. 1 vidíme graf Z-funkce. X je původní hodnota, *Φ1* a *Φ2* jsou horní a dolní meze.

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

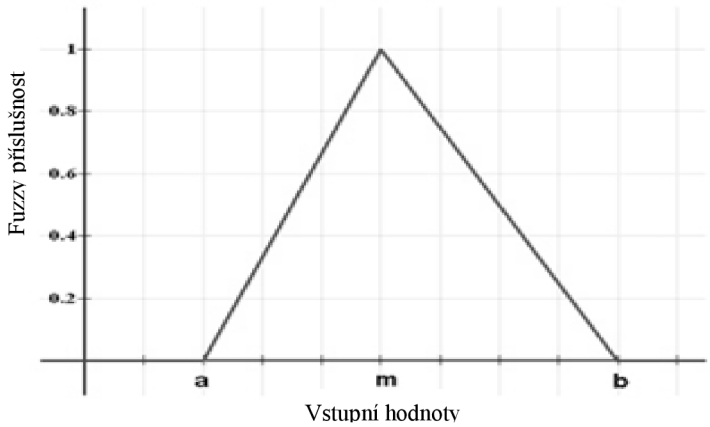


Obr. 1: Graf Z-funkce

Trojúhelníková funkce [1]

Definována dolní mez *a*, horní mez *b* a hodnotou m, kde $a < m < b$. Její graf vidíme na Obr. 2.

$$\mu_A(x) = \begin{cases} 0, & x \le a \\ \frac{x-a}{m-a}, & a < x \le m \\ \frac{b-x}{b-m}, & m < x < b \\ 0, & x \ge b \end{cases} \quad (2)$$

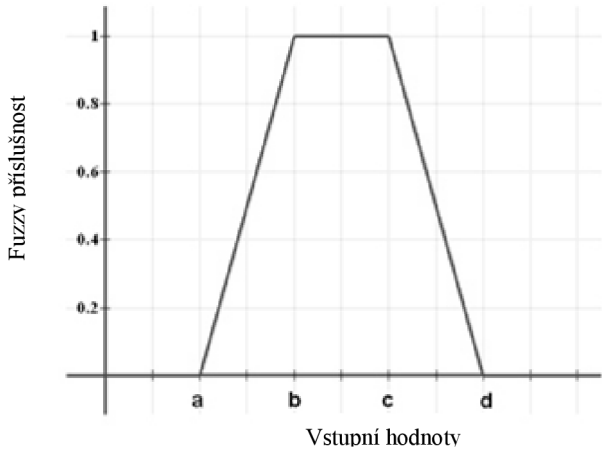


Obr. 2: Graf trojúhelníkové funkce

Lichoběžníková funkce [1]

Je definována dolní mez *a*, horní mez *d*, dolní hranice podpory *b*, a horní hranice podpory *c*, kde $a < b < c < d$. Její graf je na Obr. 3.

$$\mu_A(x) = \begin{cases} 0, & (x < a) \text{ or } (x > d) \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & b \le x \le c \\ \frac{d-x}{d-c}, & c \le x \le d \end{cases} \tag{3}$$



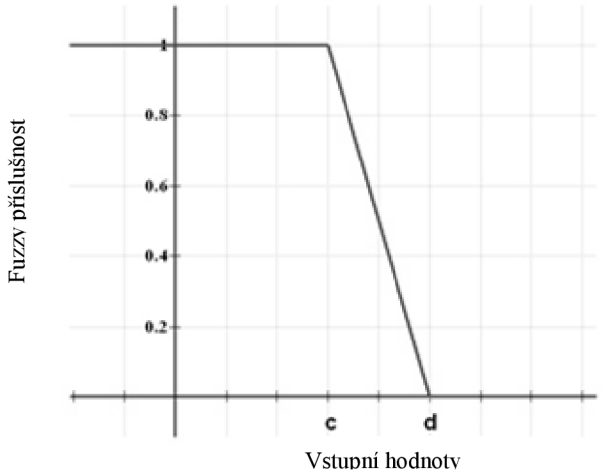
Obr. 3: Graf lichoběžníkové funkce

Dva speciální případy lichoběžníkové funkce: [1]

R-funkce

Má parametry $a = b = -\infty$, horní mez d a horní hranice podpory c. Graf je na Obr. 4.

$$\mu_A(x) = \begin{cases} 0, & x > d \\ \frac{d-x}{d-c}, & c \le x \le d \\ 1, & x < c \end{cases} \tag{4}$$

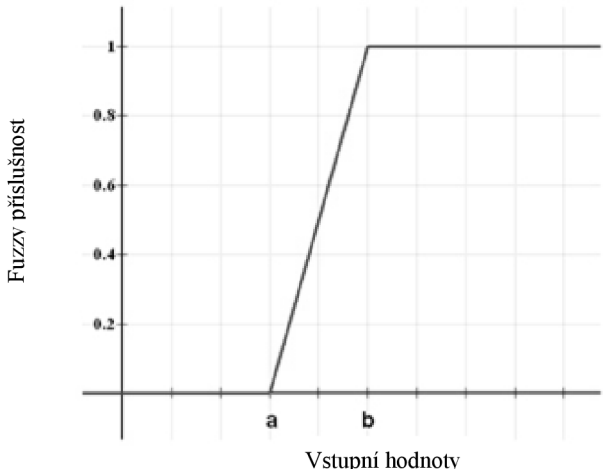


Obr. 4:Graf Rfunkce

L-Funkce

Má parametry $c = d = -\infty$, dolní mez a a dolní hranice podpory b. Graf je na Obr. 5.

$$\mu_A(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases} \tag{5}$$

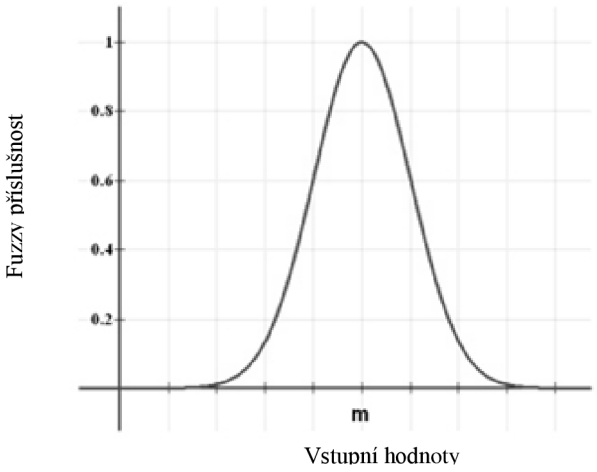


Obr. 5: Graf L funkce

Gaussova funkce [1]

Definována střední hodnotou m a směrodatnou odchylkou $k>0$. Čím menší je k, tím je funkce užší. Její graf je na Obr. 6.

$$\mu_A(x) = e^{-\frac{(x-m)^2}{2k^2}} \tag{6}$$



Obr. 6: Graf gaussovy funkce

1.1.3. Fuzzy množinové operace

Fuzzy množiny jsou rozšířením klasických množin: mají-li prvky fuzzy množiny své funkce příslušnosti definovány skokově (hodnoty pouze 0 nebo 1), pak **fuzzy množinové operace** dají stejné výsledky jako klasické množinové operace. Obvyklé definice základních množinových operací pro fuzzy množiny (označené zde jako ***A***, ***B***, ***C***) a prvek x z univerza X jsou následující [$A(x)$, $B(x)$, resp. $C(x)$ znamenají hodnoty příslušnosti x do ***A***, ***B,*** resp. ***C***]: [17]

- sjednocení: $C(x) = \max [A(x), B(x)]$ pro $\boldsymbol{C} = \boldsymbol{A} \cup B$,
- průnik: $C(x) = \min [A(x), B(x)]$ pro $\boldsymbol{C} = \boldsymbol{A} \cap B$,
- doplněk: $\neg A(x) = 1 - A(x)$

1.2. Fuzzy C-means

Patří mezi nehierarchické metody shlukování (rozkládají danou množinu do podmnožin dle předem daného kritéria).

1.2.1. C-means clustering

Vezme se N sekvencí takových, že $S=\{C\}^I$, kde C={A, C, G, T}. Náhodně se vybere K sekvencí jako základní soubor (K musí být menší než N). Poté se spočítají nukleotidové frekvence nebo obsah GC pro všechny sekvence. Tyto výpočty se pak použijí pro klasifikaci. Sekvence je přiřazena do třídy, se kterou má největší fuzzy podobnost. Ta se vypočítá tak, že $x_1,\ldots,x_P$ je soubor reálných čísel z P. Počet iterací je dán parametrem r. Fuzzy vážený průměr (weighted fuzzy average (WFA)) se vypočítá pomocí následující rovnice:

$$\mu^r = \sum_{p=1}^{P} w_p{}^{(r)} x_p \,, \; r = 0,1,2 \ldots \tag{7}$$

X je parametr nebo příznak a p je číslo parametrů. Vzdálenost $d_{i,j}$ kde i=0 … N, a j=0 … k, se vypočítá: $D_{i,j} = \max(\mu^r_j)$, pro všechny j. Počáteční průměr a Gaussian je vycentrován nad průměrem a váha w_p je spočítána pro x_p.
Příznakové vektory jsou přiděleny ke každému členu základního souboru. Prázdné nebo malé třídy se neberou v potaz. Třídy, které jsou si blízké, jsou zahrnuty do jedné třídy. Cluster centra jsou nahrazeny váženými Fuzzy průměry a jsou přiřazeny příznakové vektory. To se opakuje, dokud nedojde ke konvergenci. [36], [20]

1.2.2. Aplikace pro fuzzy

Ve fuzzy shlukování se počítá pravděpodobnost příslušnosti do určitého shluku pro každý bod. Rozložení objektů ve shlucích je pak popsáno tak, že body na okraji mají nižší stupeň příslušnosti než body v centru. Objekt tak může patřit do více shluků zároveň. Lépe se pak identifikují objekty, které nelze přiřadit do žádného shluku.
V případě, že každý objekt má pravděpodobnost příslušnosti k nějakému shluku rovnu jedné a k ostatním nulovou, pak je výsledkem pevné shlukování. Naopak jednotlivé shluky jsou neurčitelné, pokud se stupeň příslušnosti každého objektu k libovolnému shluku rovná převrácené hodnotě počtu shluků.
Součet příslušností jednotlivých objektů ke všem shlukům je 1.
Algoritmus pro fuzzy C-means je následující:

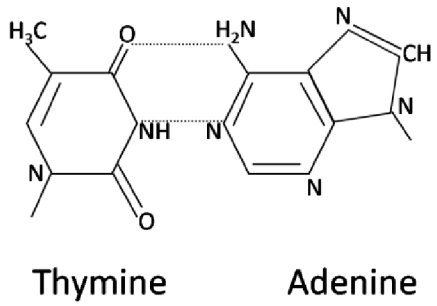
1. Vybereme počet shluků.
2. Náhodně přiřadíme každému bodu koeficient příslušnosti.
3. Spočítáme střed každého shluku.
4. Pro každý bod spočítáme pravděpodobnost příslušnosti k určitému shluku.
5. Opakujeme krok 3 a 4, dokud změna koeficientu příslušnosti není menší, než daný práh citlivosti.

Algoritmus minimalizuje rozdíly uvnitř shluku, ale stejně jako u K-means algoritmu se jedná o lokální minima a výsledek závisí na zvolených mírách příslušnosti. [13]

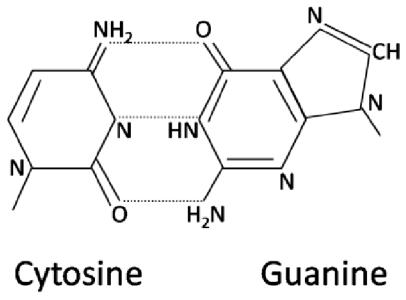
2. Klasifikační kritéria

2.1. GC obsah

DNA obsahuje 4 druhy nukleotidů (A, C, G a T), které mají mezi sebou vodíkové vazby. Adenin se páruje s thyminem a mají dva vodíkové můstky, jak je znázorněno na Obr. 7. Guanin se páruje s cytosinem (Obr. 8) a jsou mezi nimi 3 vodíkové můstky. Tato vazba je tedy stabilnější.
Proto se GC podíl někdy používá ke stanovení určité charakteristiky o zkoumané sekvenci. Organismy jsou většinou popsány všemi 4 nukleotidy A, C, G a T. Toho je využíváno pro rozdělení genomu organismů. Některé ovšem obsahují vyšší procento GC a jsou tak známé jako GC-rich, zatímco jiné organismy mají vyšší zastoupení AT a jsou známé jako AT-rich. [17], [5]



Obr. 7: Thymin a adenin



Obr. 8: Cytosin a guanin

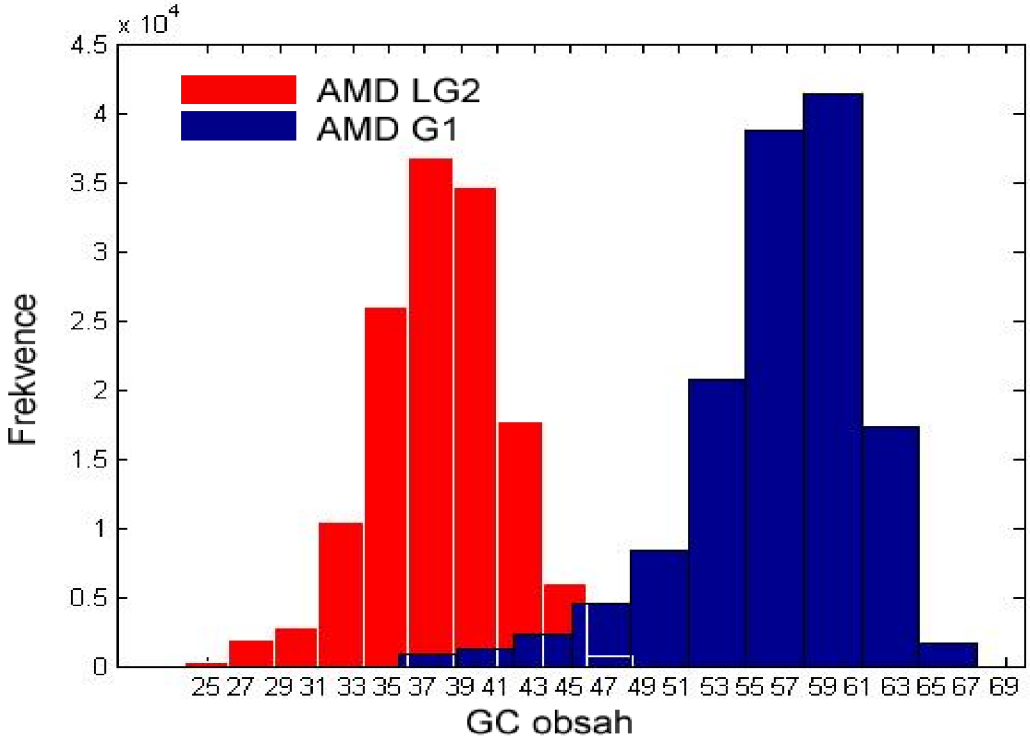
Organismy patřící do kmene Actinobacteria jsou známé jako GC rich. Na druhé straně, Arabidopsis thaliana, často zkoumaný organismus, obsahuje méně než 40 % CG, z čehož plyne, že je bohatší na AT. [17]

$$\frac{C+G}{A+C+T+G} \times 100\% \tag{8}$$

V rovnici značí A, C, G, T počty jednotlivých nukleotidů.

Příklad pro GC obsah

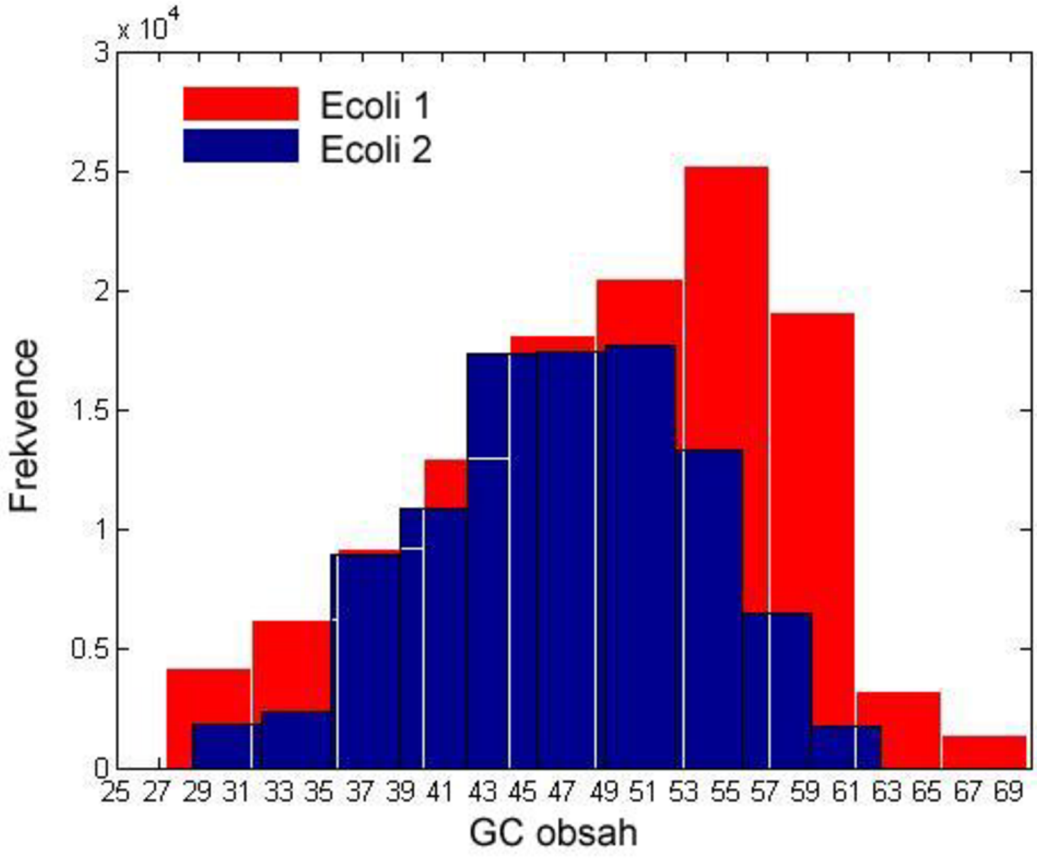
Vzorky byly projížděny oknem o velikosti 700 nukleotidů, čímž byly „rozděleny na jednotlivé fragmenty“ a pro každou pozici okna spočítán GC podíl. Ten byl pak zobrazen v histogramu, který sdružil pozice okna se stejným GC podílem. Rozsah hodnot je stanoven na 25 – 75 %, protože se obvykle podíl nepohybuje pod nebo nad těmito hranicemi.
Nejprve byly analyzovány 2 třídy patřící do Archea a Bacteria které jsou fylogeneticky vzdálené. Z histogramů těchto dvou vzorků je vidět, že jejich podíly GC se v jednotlivých fragmentech odlišují, překryv je jen v malé oblasti.



Obr. 9: Histogram obsahu GC pro délku okna 700 - AMD LG2 je Leptospirillum sp. group II a AMD G1 je Ferroplasma sp. Typ II .

Na Obr. 9 vidíme analýzu sekvence Leptospirillum (gi|206601869|gb|DS995261.1| Leptospirillum sp. Group II), jejíž délka je 138 038 bp a sekvence Ferroplasma (gi|42538182|gb|CH004071.1| Ferroplasma sp. Type II), jejíž délka sekvence je 489 309 bp (analyzována jen do délky první sekvence).

Další analýza je provedena na vzorcích fylogeneticky blízkých, a to dvou Ecoli. Histogramy na Obr. 10 se identicky překrývají, protože oba tyto vzorky patří do stejné skupiny bakterií, a jejich obsah GC je velmi podobný. Prvním je Ecoli (gi|83404799|ref|NC_007635.1| Escherichia coli plasmid pCoo) s délkou sekvence 98 396 bp, druhou Ecoli (gi|116006783|ref|NC_008460.1| Escherichia coli plasmid pO86A1) s délkou sekvence 120 730 bp.



Obr. 10: Histogram obsahu GC pro délku okna 700 u dvou Ecoli.

GC obsah se může někdy ukázat jako vhodný parametr pro samostatné fragmenty. Nicméně, je potřeba vzít v úvahu některé faktory, například, že obsah GC má větší význam v kódujících oblastech genu. [17]

2.2. Nukleotidové frekvence

Další možností pro popis genomových sekvencí je frekvence jednotlivých nukleotidů nebo oligonukleotidů. Oligonukleotidy jsou krátké sekvence, s pevně danou délkou, obvykle do 20 bází. Mezi oligonukleotidy patří například dinukleotidy, trinukleotidy či tetranukleotidy. Příkladem dinukleotidů jsou AA, AG, AC, AT, CC, CG, CT, CA, TT, TG, TC, TA, GA, GT, GC, GG, tzn. 4^2=16 možností. Trinukleotidů s příkladem ACG už je podstatně více a to 4^3=64. Tetranukleotidů je 4^4=256. [31]

2.3. Sekvenční motivy

Jsou to krátké sekvence nukleotidů nebo aminokyselin, které se vyskytují v mnoha sekvencích a které mají nebo mohou mít určitý biologický význam.

Často v sekvenci slouží pro označení specifických vazebných míst pro proteiny, jako jsou nukleázy nebo transkripční faktory (TF). Jiné jsou zapojeny do důležitých procesů na úrovni RNA, včetně vazby ribosomů a zpracování mRNA (sestřih, editace, polyadenylace) a ukončení transkripce.

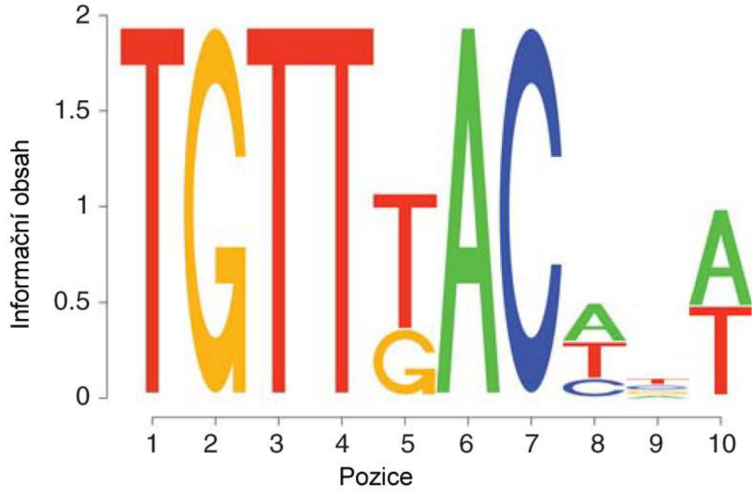
Pokud mluvíme o motivu u proteinů, jedná se o strukturální rysy, které jsou tvořeny trojrozměrným uspořádáním aminokyselin, jež spolu nesousedí. DNA motivy také poskytují signál pro vazbu bílkovin nebo vznik záhybů.

Existuje mnoho strukturálních prvků (motivů), které jsou zachovány u různých proteinů. Sacharidy se v bílkovinách připojují na aminokyselinu asparagin prostřednictvím N-glycosylation míst. N-glykosylační motiv, který se vyznačuje konsensuální sekvencí Asn-Xaa-Ser/Thr, kde první aminokyselinou je asparagin (Asn), druhá aminokyselina může být jakákoliv z 20 aminokyselin (XAA) a třetí aminokyselina je buď serin (Ser) nebo threonin (Thr).

Nicméně to, že se tento motiv v sekvenci objeví, ještě neznamená, že je v daném místě glykosylována.

Můžeme se také podívat na složitější motivy nebo domény, jako jsou aktivní místa enzymů nebo receptorů vazebných míst. [4], [26], [2], [25]

2.3.1. Sekvenční logo



Obr. 11: Sekvenční logo - konsensus sekvence FOXA3 transcripčního faktoru [22]

Sekvenční logo je grafické znázornění stálosti sekvence nukleotidů u DNA či RNA nebo aminokyselin u proteinových sekvencí. Je vytvořeno na základě zarovnaných sekvencí a zobrazuje jejich konsensus a zároveň rozdílnost. Můžeme ho vidět na Obr. 11.

Sekvenční loga jsou často používána pro znázornění vlastností dané sekvence, jako jsou například protein-vazebné místa v DNA nebo funkční celky v proteinech. Konsensuální logo je zjednodušená variace sekvenčního loga. Stejně jako sekvenční logo je vytvořeno ze zarovnaných proteinů, DNA nebo RNA sekvencí a dává nám informace o konzervovanosti každé pozice v sekvenčním motivu nebo v zarovnaných sekvencích. Konsensuální logo zobrazuje pouze míru zachování informace a ne frekvenci jednotlivých nukleotidů nebo aminokyselin v každé pozici. Místo toho zobrazuje relativní četnost každého znaku, tzn. míru zachování každé pozice pomocí výšky znaku na dané pozici.

Celková výška písmen znázorňuje informační obsah pozice v bitech. Pro DNA: vycházíme z předpokladu, že k zakódování jedné báze je potřeba dvou bitů. [26], [2]

2.4. DNA Barcoding

DNA Barcoding je technika, umožňující identifikaci a klasifikaci neznámých vzorků pomocí krátkého úseku sekvence DNA. U živočichů a mnoha eukaryot se jako nejvhodnější ukázala být pro barcoding mitochondriální DNA a její gen CO1. Tento gen vyrábí klíčový enzym „podjednotka 1 cytochrom oxidáza“, který je součástí dýchacího řetězce, kde katalyzuje redukci kyslíku na vodu.

Tento gen obsahuje standardně 648 párů bází. Mitochondriální DNA byla vybrána také proto, že jediná buňka obsahuje až 1 000 mitochondrií, ve kterých je více kopií DNA. Proto i malý vzorek může posloužit k získání dostatečného množství DNA pro úspěšné sekvenování.
Je vhodný, protože v rámci druhu se tento úsek liší jen nepatrně, naproti tomu rozdíly mezi druhy jsou patrné. Například při porovnání člověka a šimpanze se tento gen liší asi na 60 místech. Pro spolehlivou detekci je nutné popsat alespoň 10 jedinců každého druhu. Klasický morfologický popis druhu potřebuje více exemplářů obou pohlaví, zatímco pomocí barcodingu analyzujeme i malou část těla.
Navíc 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.

Obr. 12: DNA Barcode

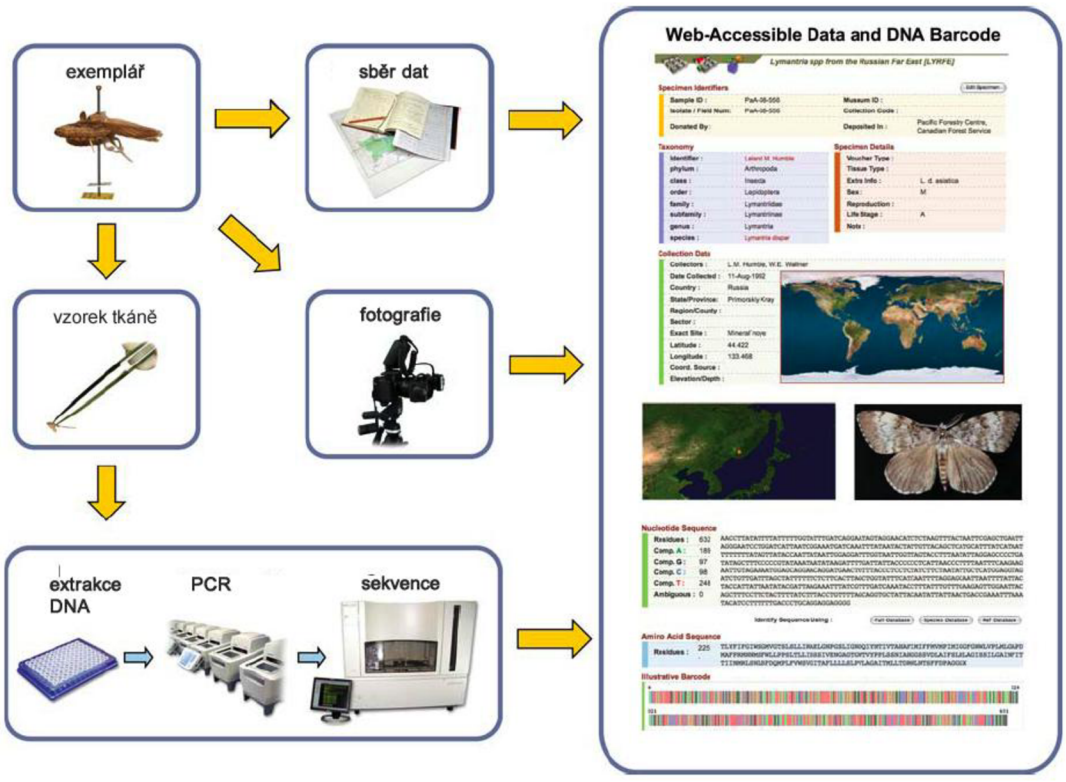
2.4.1. Postup DNA barcodingu

Postup DNA barcodingu je následující a můžeme ho vidět také na Obr. 13.

1. Vzorky z různých úložišť biologických materiálů slouží k vytvoření podkladu pro vlastní identifikaci. Mezi hlavní zdroje vzorků patří například různé sbírky, zoologické zahrady, muzea, ale i volná příroda.

2. Ze vzorků se v laboratořích pomocí PCR získá požadovaná sekvence DNA čárového kódu. Tyto údaje pak putují do databází, v kterých probíhá další zpracování. V nejmodernějších laboratořích může tato analýza trvat jen několik hodin. Je získána jeho DNA sekvence, fylogenetické informace, informace o prostředí kde žije, fotografie různých jedinců atd.
3. Databáze jsou nejdůležitější částí řetězce. Proto je snaha vytvořit centralizovanou databázi všech živočišných druhů. Tato databáze by sloužila k porovnání neznámého vzorku se záznamy v databázi a tím k určení původu vzorku.

   Největší 3 databáze DNA – GenBank, EMBL a DDBJ se pro záznam dat DNA barcodingu dohodly na datovém standardu CBOL.
4. Pomocí analýzy dat můžeme porovnávat jednotlivé záznamy v databázích. CBOL nabízí portál, který umožňuje vědcům jednoduše ukládat, spravovat, analyzovat a zobrazovat jejich barcoding záznamy.
5. Stačí část, či vzorek neznámého organismu a po získání DNA sekvence můžeme během pár sekund přes webové rozhraní databáze zjistit, o jaký druh organismu šlo, případně o něm získat detailnější informace



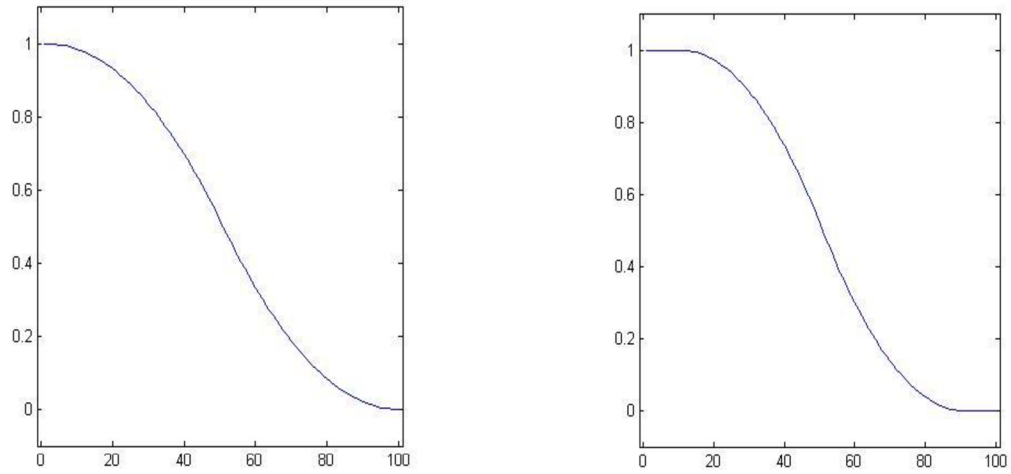
Obr. 13: Postup při DNA barcodingu

2.4.2. Fuzzy funkce pro barcoding

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.

Využívá se fuzzy z funkce příslušnosti, jejíž rovnice (1) již byla uvedena výše. X je definováno jako vzdálenost 2 sekvencí nebo sekvence od referenčních sekvencí jednotlivých druhů. Dva parametry Φ1 a Φ2 je třeba odhadnout dle konkrétního souboru dat. Jedná se o maximální vnitrodruhovou a minimální mezidruhovou vzdálenost sekvencí.

Obvykle se za Φ1 a Φ2 dosazuje 1 - 10 percentil pro vzdálenost v rámci druhu respektive 90 - 100 percentil pro mezidruhovou vzdálenost. [37], [33]



Obr. 14: Z funkce s parametry Φ1=10, Φ2=90 vlevo a s parametry Φ1=1, Φ2=99 vpravo

2.4.3. Výpočet vzdáleností sekvencí

Výpočet vzdálenosti pomocí denzity [30]

Sekvence jsou nejprve převedeny do numerického formátu, a to způsobem, že za adenin se dosadí 1, za cytosin 2, za guanin 3 a za thymin 4. Poté sekvenci projíždí okno zadané délky a počítá denzitu tak, že zprůměruje hodnoty v okně (všechny sečte a vydělí délkou okna). Výstupem je matice hodnot denzit. Denzity sekvencí jsou mezi sebou porovnány, každá s každou. Podobnost je počítána pomocí euklidovské vzdálenosti, pomocí vzorce:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{10}$$

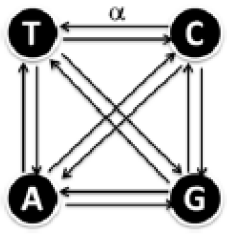
kde n je délka sekvencí, x je denzita jedné sekvence na pozici i a y je denzita druhé sekvence na pozici i.
Výstupem je vzdálenost, která je součtem vzdáleností jednotlivých částí sekvencí. [14]

Výpočet vzdálenosti pomocí Jukes-Cantor [30], [10]

Tento model určuje pravděpodobnost změny jednoho stavu systému na druhý a je vhodný pro výpočet vzdálenosti mezi dvěma sekvencemi.

$$d_{JC} = -\alpha \ln(1 - \frac{p}{\alpha}) \tag{11}$$

α je parametrem, který udává pravděpodobnost změny na jiný znak. J-C model předpokládá, že všechny typy záměn mají stejnou pravděpodobnost, grafické znázornění pro nukleotidy je vidět na následujícím Obr. 15.



Obr. 15: Jukes-Cantor [30]

Z Obr. 15 lze tedy vyvodit, že hodnota parametru α pro genomické sekvence bude α=3/4, tedy:

$$d = -\frac{3}{4}\ln(1 - \frac{4}{3}p) \tag{12}$$

3. Testovací data

3.1. Získávání genomických dat

3.1.1. Sekvenování DNA

Sekvenování DNA je vlastně zjišťování pořadí nukleových bází (A, C, G, T) v sekvencích DNA. To probíhá pomocí mnoha biochemických metod. [23], [17], [7]

Tabulka 1: Kódování DNA eukaryot

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

3.1.2. Databáze DNA

Mnoho databází, programů a analýz týkající se geonomu je přístupných veřejně na Internetu. V České republice je např. v ÚMG AV ČR organizována databáze užitečných bioinformatických linek. Velká bioinformatická centra v Evropě jako European Bioinformatic Institute a Expert Protein Analysis System nabízejí celou řadu zajímavých služeb a databází. V USA je jedním z nejvíce využívaných zdrojů informací National Center for Biotechnology Information.

EMBL

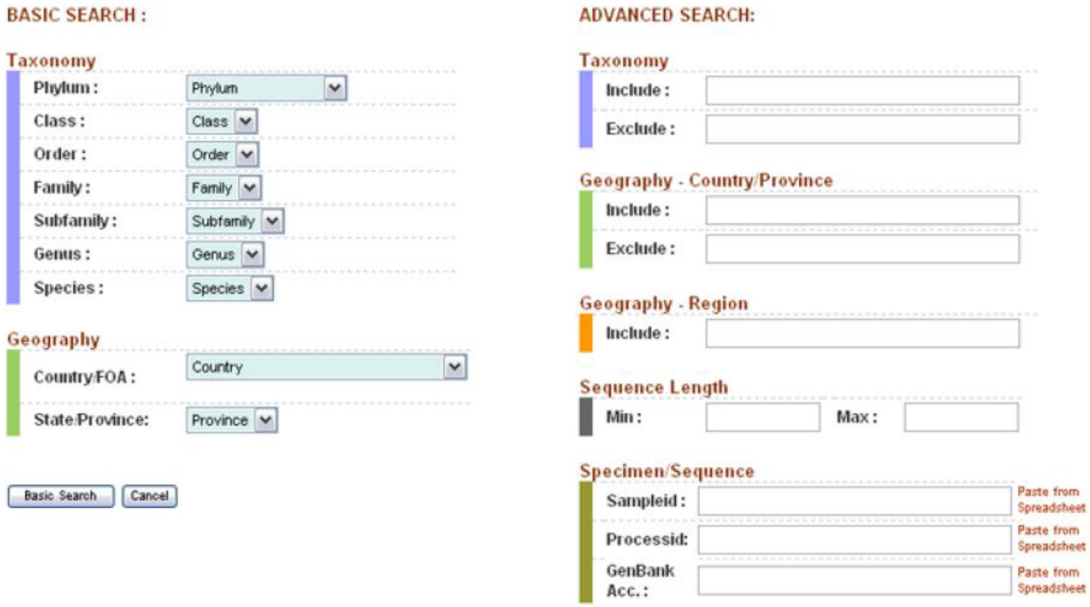
EMBL obsahuje skoro čtrnáct miliard nukleotidů, tvořících mnoho genů a genomů z různých organismů. Databáze EMBL je organizována Evropskou molekulárně biologickou laboratoří (EMBL). Je to veřejná evropská primární nukleotidová databáze se sídlem v Anglii na adrese http://www.ebi.ac.uk/embl. Databáze je vytvářena v součinnosti s ostatními nukleotidovými databázemi GENBANK (USA) a DDBJ (Japonsko) a je velmi dobře přístupná spolu s mnoha odvozenými a dalšími databázemi přes SRS (Sequence Retrieval System), například na adrese http://srs6.ebi.ac.uk. Databáze obsahuje všechna data zaslaná vědeckou komunitou, a to bez kontroly. Z tohoto důvodu může obsahovat určité procento chyb. Manuál k databázi je k dispozici na adrese http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html.

GenBank

GenBank je databáze sekvencí, kterou tvoří otevřený, anotovaný soubor všech veřejně dostupných nukleotidových sekvencí a jejich proteinových překladů. Tato databáze je organizována Národním institutem zdraví (NIH) v USA. GenBank a její spolupracovníci dostávají sekvence z laboratoří po celém světě z více než 100.000 různých organismů. GenBank nadále roste exponenciální rychlostí, zdvojnásobuje se každých 18 měsíců. V srpnu 2006, obsahovala více než 65 miliard nukleotidů a více než 61 milionů sekvencí. Do GenBank přispívají přímo jak jednotlivé laboratoře, tak hromadné podání ve velkém měřítku ze sekvenovacích center. Podrobnější informace o databázi lze najít například na adrese: http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html. [6]

BOLD

Barcode of Life Data Systems (BOLD) - databáze sloužící k ukládání, analýze a zveřejňování záznamů čárového kódu DNA. BOLD je volně dostupná pro každého výzkumníka, který se zajímá o DNA barcoding. Záznamy zde splňují normy nutné pro BARCODE data společná pro všechny databáze. BOLD má společné rozhraní pro 3 způsoby vyhledávání. A to hledání ve všech záznamech, v souboru projektů nebo v jednom konkrétním projektu.
Dále existují dva typy vyhledávání v BOLDu - základní vyhledávání a pokročilé vyhledávání. Základní vyhledávání umožňuje pomocí rolovacího menu vyhledávat podle taxonomie nebo zeměpisné polohy záznamu v BOLD. Pomocí rozevíracího výběru se obnovují volby pro další úrovně výběru. Pokročilé vyhledávání umožňuje více specifik. Tím můžeme zúžit rozsah každého hledání. V pokročilém vyhledávání můžeme také vyloučit určitá kritéria, aby se zabránilo nechtěným výsledkům. Na Obr. 16 vidíme vyhledávací stránku BOLDu.[24] V Tabulka 2 jsou pak shrnuty barcode databáze.



Obr. 16: Pokročilé vyhledávání na BOLDu [24]

Tabulka 2: Databáze barcode

Jméno databáze	Adresa	Organismy
Algaterra	www.algaterra.org	Řasy
BOLD	www.boldsystems.org	Všechny organismy
ISTH	www.isth.info	Trichoderma (Houby)
Mycobank	www.mycobank.org	Houby
Nematol	http://nematol.unh.edu/	Hlístice
Silva	www.arb-silva.de	Bacteria, Archaea, Eukarya
Sponge Barcoding Project	www.spongebarcoding.org	Mycí houby
MOTU	http://nemhelix.cap.ed.ac.uk:880	Hlístice
Unite	http://unite.ut.ee	Houby ektomykorhizní

3.1.3. Vyhledávání sekvencí v databázích

Způsobů, jak získat požadovaná data z databází, je několik. Každá sekvence má svůj unikátní identifikátor, podle kterého ji můžeme jednoduše vyhledat. Další možností je hledání podle klíčových slov. Mezi dalšími vyhledávacími kritérii mohou být druh organismu, autor, rok, místo původu a další. [28]

3.1.4. Formáty sekvenčních dat

Společně s rozvojem bioinformatiky se vyvinula široká řada formátů uchovávaných dat. Některé z nich se uchytily a našly široké uplatnění, jiné nikoliv. Obecně závisí úspěch formátu na jeho použitelnosti v různých případech a souvislostech, tedy na způsobu a množství uchované biologické informace a její přehlednosti a jednoduché dostupnosti uživateli, či softwaru pro analýzu biologických dat. Mezi nejznámější formáty patří: RAW DATA, FASTA, PIR, EMBL, SWISSPROT, IG, GenBank flatfile. [6]

FASTA

V dnešní době je nejpoužívanějším formátem pro práci s biologickými daty. Hodí se jak pro práci s nukleotidovými, tak i aminokyselinovými sekvencemi. Není příliš vhodný pro archivaci. Na rozdíl od vnitřních formátů databází neumožňuje uložit dodatečné informace pro databázové vyhledávání. To je na druhou stranu jeho výhoda při práci, jelikož se dá zapsat přímo z klávesnice. Soubor ve formátu FASTA je textovým souborem, který začíná na prvním řádku znakem > (větší než). Za tímto znakem následuje „hlavička“, ve které je obsažen název sekvence, anotace a různé další údaje, které nejsou obsažené ve vlastní sekvenci, jako například zdrojová databáze apod. Nejdůležitější částí hlavičky je identifikátor, který je reprezentován skupinou alfanumerických znaků hned za znakem >. Tento identifikátor je jedinečný a musí být v hlavičce zahrnut, další informace lze už považovat za volitelné. Za hlavičkou pak následuje vlastní sekvence ve formě surových dat, ta by neměla obsahovat mezery či prázdné řádky. Velkou výhodou FASTA formátu je možnost spojit do jednoho souboru více sekvencí, které pak mohou být zpracovány naráz. Podmínkou ale je, aby byly všechny sekvence buď nukleotidové, anebo aminokyselinové, jelikož FASTA neumožňuje přímo specifikovat typ sekvence. Nepřímo, což některé programy vyžadují, je to možné pomocí koncovky textového souboru, kde `*.nt` označuje nukleotidovou sekvenci a `*.aa` sekvenci aminokyselin. Dalšími koncovkami mohou být např: `*.fa`, `*.fas`, `*.fasta`, `*.fsa` a další, neboť neexistuje žádný standard. V Tabulka 3 jsou vypsány identifikátory hlavičky FASTA pro nejpoužívanější databáze. [28], [6]

Tabulka 3: Identifikátory v hlavičce FASTA pro nejpoužívanější databáze

GenBank	gi|gi-number|gb|accession|locus
EMBL Data Library	gi|gi-number|emb|accession|locus
DDBJ, DNA Database of Japan	gi|gi-number|dbj|accession|locus
NBRF PIR	pir||entry
Protein Research Foundation	prf||name
SWISS-PROT	sp|accession|name

3.2. Hemoglobin

Je to červený transportní metaloprotein, který přenáší kyslík v červených krvinkách u obratlovců a některých dalších živočichů. Jeho hlavní funkcí je transport kyslíku z plic do tkání a v opačném směru transport oxidu uhličitého. Jeho molekula se skládá z bílkovinné složky globinu a z prostetické skupiny hemu (pigment s obsahem železa). Globin je tvořen 4 polypeptidovými řetězci. V aktivních erytrocytech savců hemoglobin tvoří 35 % obsahu. Průměrné množství hemoglobinu v jednom erytrocytu je 28-32 pg.
U člověka se během vývoje vyskytují různé typy hemoglobinu, mají stejný hem, liší se ale v bílkovinné složce.

Hemoglobin dospělého typu (HbA)

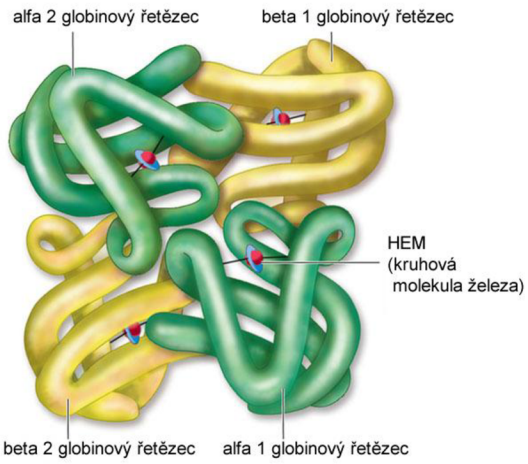
Dospělý člověk ho má v erytrocytech přibližně 98 %. HbA obsahuje 2 řetězce alfa (α), každý se skládá ze 141 aminokyselin, a 2 řetězce beta (ß), každý se skládá ze 146 aminokyselin. Je tvořen z větší části HbA_0 (neglykovaný) a asi z 5 % HbA_1 (glykovaný).

Hemoglobin dospělého typu (HbA2)

Také syntetizován v dospělosti. Místo ß podjednotek obsahuje dvě δ podjednotky. Podílí se 2,5 % na celkovém hemoglobinu.

Fetální hemoglobin (HbF)

Syntetizován ve větším množství u plodu (u dospělých jen 0,5 %). Místo ß podjednotek má dvě γ podjednotky. Má vyšší afinitu ke kyslíku.



Obr. 17: Molekulová struktura hemoglobinu [21]

Molekulová struktura hemoglobinu je zobrazena na Obr. 17. Každý monomer globinu má **primární strukturu**, která je určena pořadím aminokyselin. Dále je uspořádán do **sekundární struktury** osmi α-helixů, označované písmeny A - H. Podle polarity aminokyselin v globinu se molekula ve vodě uspořádá do **terciární struktury** (hydrofobní interakce). **Kvartérní struktura** určuje prostorové uspořádání jednotlivých podjednotek tetrameru hemoglobinu a interakce mezi nimi (hydrofobní a iontové interakce). [16], [29], [12]

3.2.1. Geny pro globinové řetězce

Skupina genů příbuzných alfa genu

Nachází se na 16. chromozomu. Lokus pro alfa globin je tetraplikován: geny $alfa_1$, $alfa_2$ a 2 pseudogeny – nefunkční kopie $alfa_1$ a $alfa_2$ genu. Gen pro zeta globin je duplikován: zeta + pseudogen zeta.

Skupina genů příbuzných beta genu

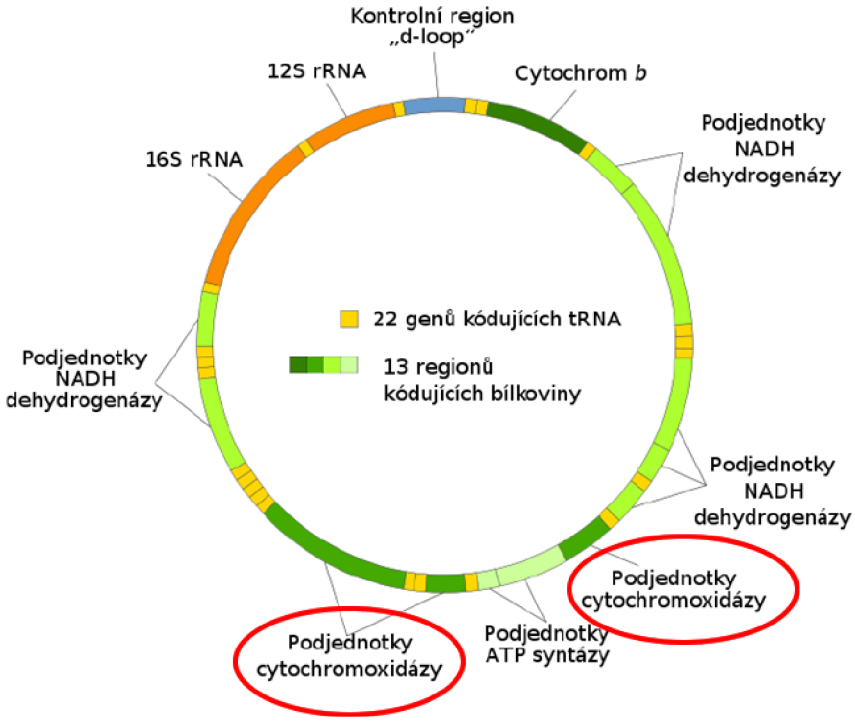
Nachází se na 11. chromozomu. Jsou to gen beta, pseudogen beta, gen delta, gen gama G, gen gama A, gen epsilon.

Mutace

Mutace beta genu postihují u heterozygotů 50 % řetězců hemoglobinu (protože je jen jeden gen beta), mutace alfa genu postihují jen 25 % molekul hemoglobinu (protože jsou 2 kopie genu alfa). Protože řetězce alfa i beta jsou kódované geny na různých chromozomech, mutace poškozují buď jen jeden, nebo druhý řetězec, nikdy ne oba současně. [16], [29], [12]

3.3. Cytochrom C oxidáza

Geny pro cytochrom C oxidázu jsou součástí mitochondriální DNA. Její strukturu můžeme vidět na Obr. 18.
Je to velký transmembránový komplex proteinů, který se jako poslední účastní elektronového transportu v mitochondriálním dýchacím řetězci. Tento enzym katalyzuje redukci kyslíku na vodu a dopravuje protony (protonová pumpa) přes biologickou membránu.
Primární struktura se skládá z řetězce asi 100 aminokyselin a molekulová hmotnost je asi 12 000 daltonů. Mnoho organismů vyššího řádu má řetězec 104 aminokyselin. Součástí komplexu proteinů u savců je integrální membránový protein složený z několika „kovových“ míst a 13 bílkovinných podjednotek. Podjednotky I - II jsou syntetizované mitochondrií a zbylé podjednotky IV - XIII jsou kódované jádrem.



Obr. 18: Schéma mitochondriální DNA člověka

Funkce

Elektrony, které pocházejí ze živin (citrátový cyklus, oxidace mastných kyselin a anaerobní glykolýza), předá kyslíku přes koenzym Q a cytochrom c. Redukuje ho 4 elektrony, přičemž vzniknou 2 molekuly vody. Přitom uvolní energii pro syntézu adenozin-5'-trifosfátu (ATP). Jako akceptor elektronů v citrátovém cyklu slouží NADH, který se oxiduje na NAD+. Meziprodukty při redukci kyslíku jsou volné radikály - peroxid vodíku a superoxid, ty však zůstávají navázané na enzym. Zjednodušeně:

$$4\ cytochrom\ C\ (red.) + O_2 + 4\ H+ \rightarrow\ 4\ cytochrom\ C\ (oxid.) + 2\ H_2O$$

Cytochrom c se také podílí na zahájení apoptózy (řízené formy buněčné smrti, používané k usmrcení buňky v procesu vývoje, v reakci na infekci nebo poškození DNA). Uvolnění velkého množství cytochromu c do cytoplasmy vede k aktivaci proteáz, které jsou zodpovědné za zničení buňky.

Využití

Cytochrom c je využíván při LLLT (Low-laserová terapie). Laserový paprsek s vlnovou délkou blížící se infračervené oblasti proniká tkání, kde v buňkách zvyšuje aktivitu cytochromu c, čímž se zvyšuje metabolická aktivita a uvolní se více energie potřebné pro buňku k regeneraci.

4. Fuzzy klasifikace struktury DNA sekvencí

Z veřejně přístupné databáze NCBI (ncbi.nlm.nih.gov) bylo vybráno 10 sekvencí, které kódují hemoglobin, konkrétně beta-globin. [19] Tento kompletní soubor dat s použitými sekvencemi je uveden v příloze.

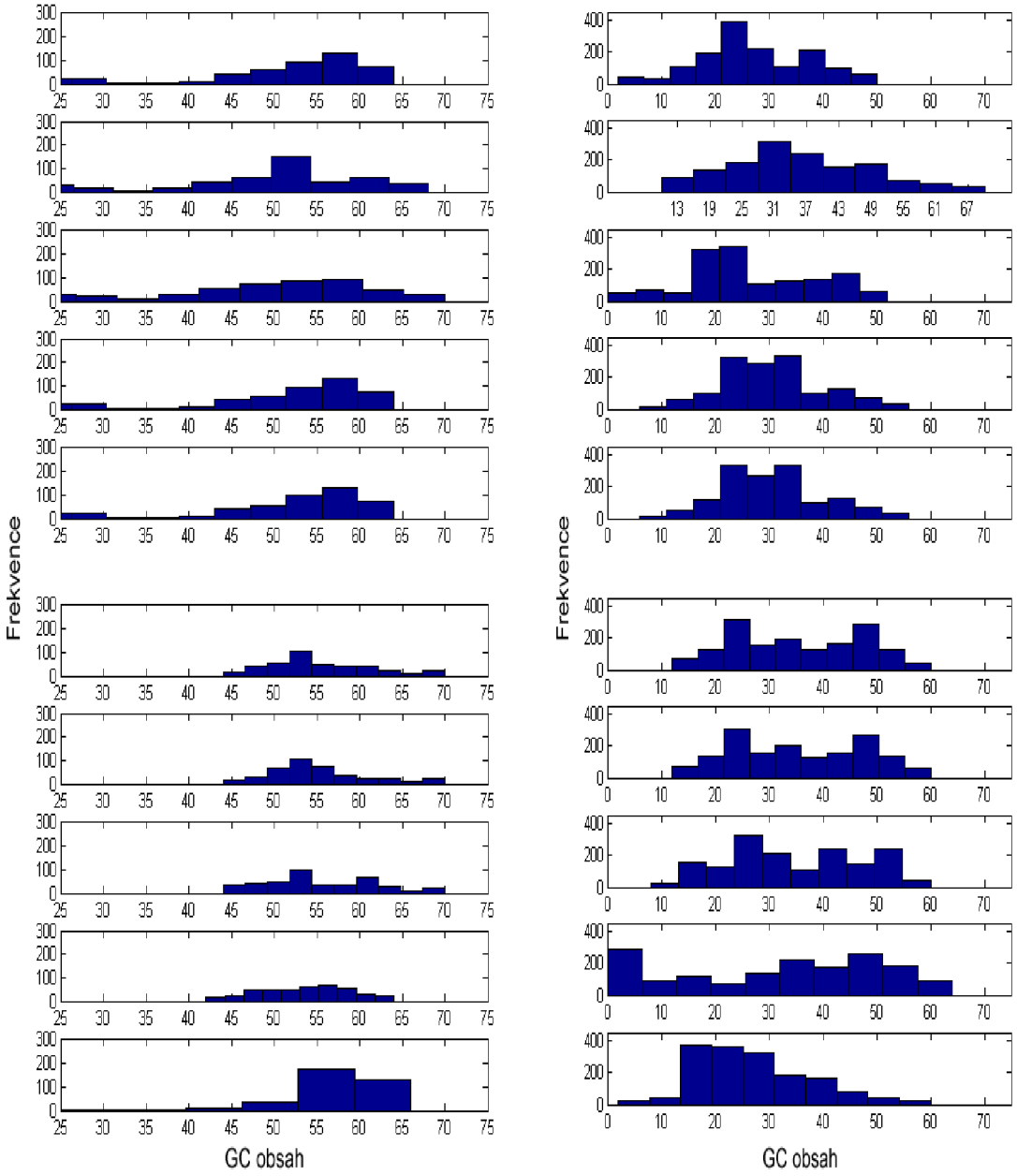
Tabulka 4: Použité sekvence

Použité sekvence									
Homo sampiens	**Mus musculus**	**Mus musculus**	**Homo sampiens**	**Homo sampiens**	**Mus musculus**	**Mus musculus**	**Mus spretus**	**Bos taurus**	**Pan trogl.**

Každá z těchto sekvencí byla rozdělena na kódující a nekódující úseky. Exony o délce 499 bp byly zarovnány. Introny měly různou délku, proto byly zkráceny na přibližně stejné úseky a poté také zarovnány s délkou 1780 bp.
U intronů i exonů byl nejprve spočítán GC podíl (metoda popsána výše), poté frekvence dinukleotidů, a trinukleotidů (jednotlivých kodonů).

4.1. GC podíl

Analýza pomocí GC podílu byla popsána již v kapitole 2.1. Vzorky byly projížděny oknem o velikosti 50 nukleotidů, čímž byly „rozděleny na jednotlivé fragmenty“ a pro každou pozici okna spočítán GC podíl. Ten byl pak zobrazen v histogramu, který sdružil pozice okna se stejným GC podílem. Výsledky pro jednotlivé sekvence, jak v kódujících tak v nekódujících úsecích, můžeme vidět na následujících histogramech.



Obr. 19: Histogram obsahu GC v kódujících (vlevo) a nekódujících úsecích (vpravo)

Z grafů na Obr. 19 je zřejmé, že obsahy GC se značně liší. A to jak pro jednotlivé sekvence, tak pro kódující a nekódující úseky. Průměrný obsah GC v intronech je 28,18, zatímco průměrný obsah v exonech je 50,16. Navržená hranice rozlišení exonů a intronů by tedy mohla být na hodnotě průměru těchto dvou hodnot a to 39,2.

4.2. Frekvence dinukleotidů

Byly spočítány počty výskytu jednotlivých dinukleotidů, jejichž číselné označení je uvedeno v tabulce 2.

Tabulka 5: Číselné označení dinukleotidů

Číslo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Dinukleotid	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT

Z důvodu rozdílné délky exonů a intronů byl počet jednotlivých dinukleotidů vydělen součtem všech nalezených dinukleotidů pro každou sekvenci zvlášť. Tím byl získán jejich poměrný podíl výskytu, viz. Tabulka 6 a Tabulka 7.

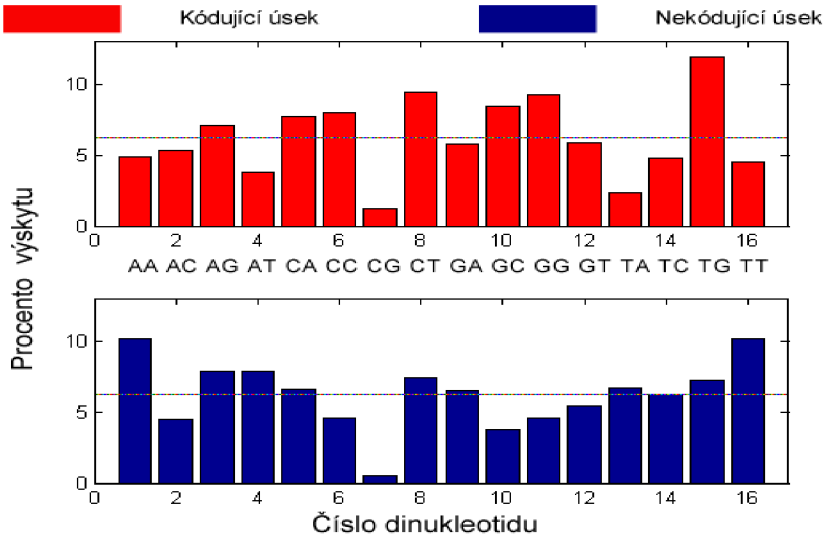
Tabulka 6: Příklad pro kódující úsek první sekvence

Dinukleotid	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
Počet	17	32	43	14	57	42	5	27	21	34	34	23	9	25	30	23
Poměrný výskyt	3,9	7,3	9,9	3,2	13,1	9,6	1,1	6,2	4,8	7,8	7,8	5,3	2,1	5,7	6,9	5,3

Tabulka 7: Příklad pro nekódující úsek první sekvence

Dinukleotid	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
Počet	175	52	91	114	72	51	6	52	74	37	35	54	101	41	67	94
Poměrný výskyt	15,7	4,7	8,2	10,2	6,5	4,6	0,5	4,7	6,6	3,3	3,1	4,8	9,1	3,7	6,0	8,4

Tyto poměrné podíly byly pro všech 10 sekvencí vyneseny do sloupcového grafu odděleně pro kódující a nekódující úseky, viz. Příloha 1 a Příloha 2. Z těchto výskytů byl pro kódující i nekódující úsek spočítán průměr, který je vynesen také ve sloupcovém grafu na Obr. 20.



Obr. 20: Průměrný výskyt dinukleotidů v jednotlivých úsecích

Z průměru je zřejmé, že se exony a introny na mnoha místech liší. Na Obr. 20 vidíme také vyznačenou hranici průměrného tj. 6,25 procentního výskytu jednotlivých dinukleotidů (100 % děleno 16 dinukleotidy = 6,25), podle které by bylo možné rozdělit sekvence na kódující a nekódující. Např. u nukleotidů AA, AT a TT jsou hodnoty u nekódujícího úseku nad touto hranicí a u kódujícího pod touto hranicí. U nukleotidů CC, GC a GG je tomu přesně naopak.
V následujících tabulkách Tabulka 8 a Tabulka 9 jsou pak červeně vyznačeny nadprůměrné hodnoty (větší než 6,25) u každé sekvence. Dále je zde také spočítán průměr pro jednotlivé dinukleotidy a také průměrná odchylka.

Tabulka 8: Poměrný výskyt dinukleotidů v kódujících úsecích

Procento dinukleotidu	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
Seq 1	3,9	7,3	9,9	3,2	13,1	9,6	1,1	6,2	4,8	7,8	7,8	5,3	2,1	5,7	6,9	5,3
Seq 2	4,8	5,3	6,4	4,1	6,4	8,0	1,1	10,8	6,4	8,9	8,9	5,0	2,5	4,3	12,8	4,1
Seq 3	3,9	5,0	10,3	4,4	12,8	8,9	1,4	6,2	3,9	8,9	8,3	5,0	2,5	6,7	6,4	5,3
Seq 4	5,3	5,5	6,2	3,0	6,6	7,8	1,1	10,0	5,5	8,0	9,8	7,3	2,1	4,6	13,5	3,9
Seq 5	5,3	5,5	6,2	3,0	6,6	7,8	1,1	10,0	5,5	8,0	9,8	7,3	2,1	4,6	13,5	3,9
Seq 6	5,0	5,3	6,2	4,3	6,6	8,2	1,1	10,5	6,2	8,7	9,1	5,0	2,5	4,3	12,8	4,1
Seq 7	4,8	5,3	6,2	4,3	6,4	8,4	1,1	10,5	6,2	8,2	9,4	5,3	2,7	4,6	12,6	4,1
Seq 8	5,3	5,0	6,2	4,3	6,2	8,0	1,4	10,5	6,4	9,1	9,1	5,0	2,5	4,1	13,0	3,9
Seq 9	5,6	3,7	7,0	4,4	5,6	6,0	1,6	9,3	6,7	8,8	9,3	6,3	2,3	3,9	13,5	6,0
Seq 10	4,9	5,4	6,2	2,7	6,2	6,7	1,3	10,2	5,9	7,5	10,5	7,5	1,6	4,9	13,7	4,6
Průměr nukleotidu	4,9	5,3	7,1	3,8	7,7	8,0	1,3	9,4	5,7	8,4	9,2	5,9	2,3	4,8	11,9	4,5
Odchylka	0,4	0,4	1,1	0,6	1,9	0,6	0,1	1,2	0,6	0,4	0,5	0,9	0,2	0,5	1,9	0,6

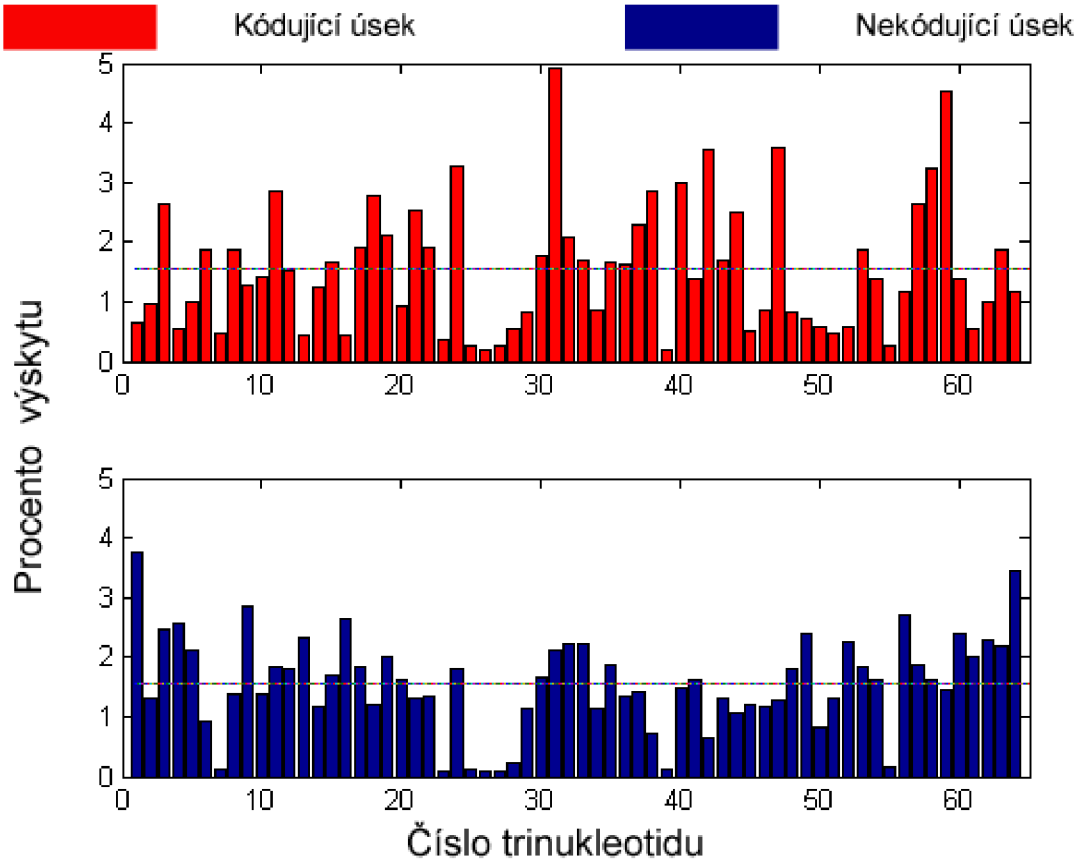
Tabulka 9: Poměrný výskyt dinukleotidů v nekódujících úsecích

Procento dinukleotidu	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
Seq 1	15,7	4,7	8,2	10,2	6,5	4,6	0,5	4,7	6,6	3,3	3,1	4,8	9,1	3,7	6,0	8,4
Seq 2	6,9	4,0	6,3	6,8	6,8	6,2	0,5	8,9	4,7	4,2	4,9	6,3	5,2	8,1	8,4	11,7
Seq 3	13,8	6,3	10,4	7,0	7,6	4,4	0,4	6,2	9,1	3,5	5,8	3,4	6,3	4,5	5,2	6,1
Seq 4	12,0	4,9	7,8	8,4	6,6	3,1	0,4	7,4	6,6	4,1	4,0	5,0	7,8	5,2	7,1	9,6
Seq 5	12,1	4,8	7,9	8,3	6,5	3,0	0,4	7,3	6,6	4,0	4,2	5,1	7,9	5,1	7,1	9,6
Seq 6	6,9	3,8	8,1	6,8	7,4	5,5	0,5	8,1	5,7	4,1	4,7	6,4	5,3	7,9	7,6	11,2
Seq 7	7,0	3,8	8,1	6,8	7,3	5,6	0,5	8,1	5,7	4,1	4,6	6,4	5,3	8,0	7,6	11,1
Seq 8	7,4	3,5	7,3	7,5	6,2	5,3	0,3	8,8	5,1	3,9	5,1	5,9	6,5	7,7	7,6	11,9
Seq 9	8,5	4,7	7,2	7,2	6,2	5,9	0,8	8,1	7,0	3,5	5,0	4,7	5,8	7,6	7,1	10,8
Seq 10	11,2	4,6	7,4	9,4	4,9	2,5	0,6	6,2	7,5	2,7	4,8	6,1	8,0	4,6	9,0	10,5
Průměr nukleotidu	10,2	4,5	7,9	7,8	6,6	4,6	0,5	7,4	6,5	3,8	4,6	5,4	6,7	6,2	7,3	10,1
Odchylka	2,5	0,5	0,6	0,9	0,5	1,0	0,1	0,9	0,8	0,3	0,5	0,7	1,1	1,5	0,7	1,2

4.3. Frekvence trinukleotidů

Obdobná analýza na stejných sekvencích byla provedena i pomocí frekvencí trinukleotidů. Na obrázcích v Příloha 3 a Příloha 4 vidíme poměrný výskyt trinukleotidů v kódujících i nekódujících úsecích pro jednotlivé sekvence.

Opět byl vypočítán průměr z jednotlivých sekvencí a ten byl vynesen do grafu s vyznačenou hranicí průměru 1,625 (vypočítáno: 100 procent děleno 64 trinukleotidy) na Obr. 21.



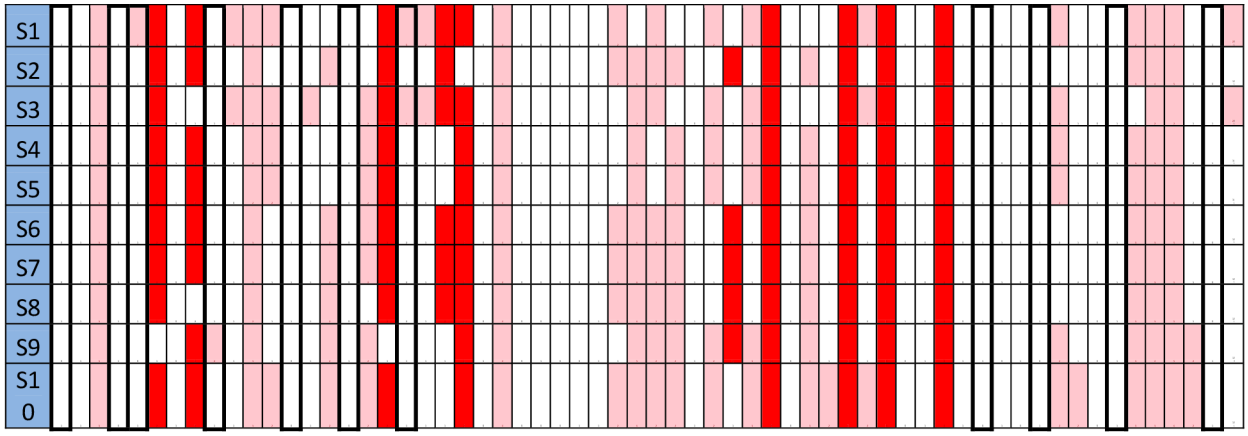
Obr. 21: Průměrný výskyt trinukleotidů v jednotlivých úsecích

Zde jsou opět zřejmé rozdíly ve výskytu trinukleotidů. Například na pozici 1, 9, 11, 20, 48, 49, 52, 54, 56, 60, 61, 62, 64 jsou nekódující úseky nad průměrem, kódující pod průměrem.

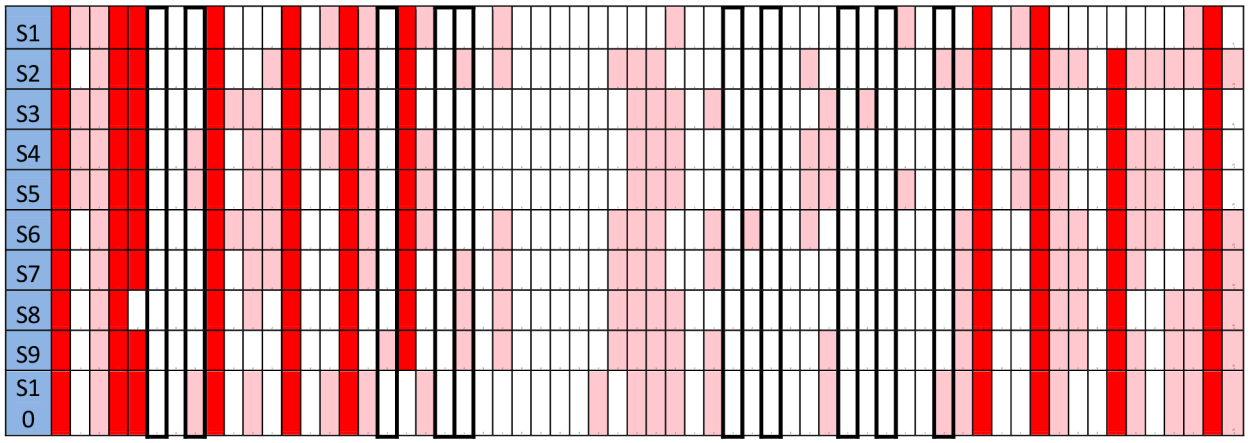
V tabulkách jsou pak světle červeně vyznačeny nadprůměrné hodnoty větší než 1,625 u každé sekvence. Z důvodu úspory místa zde nejsou uvedeny hodnoty, ale jen barevně vyznačen nadprůměrný výskyt. Sytou červenou barvou jsou navíc zvýrazněny trinukleotidy, u kterých se

hodnoty v kódující a nekódující sekvenci liší podstatně u všech 10 sekvencí (hodnota nad prahem). Analogicky nad sebou jsou u druhé sekvence podprahové hodnoty znázorněny černým ohraničením.

Tabulka 10: Nadprůměrný výskyt jednotlivých trinukleotidů v kódujících úsecích



Tabulka 11: Nadprůměrný výskyt jednotlivých trinukleotidů v nekódujících úsecích

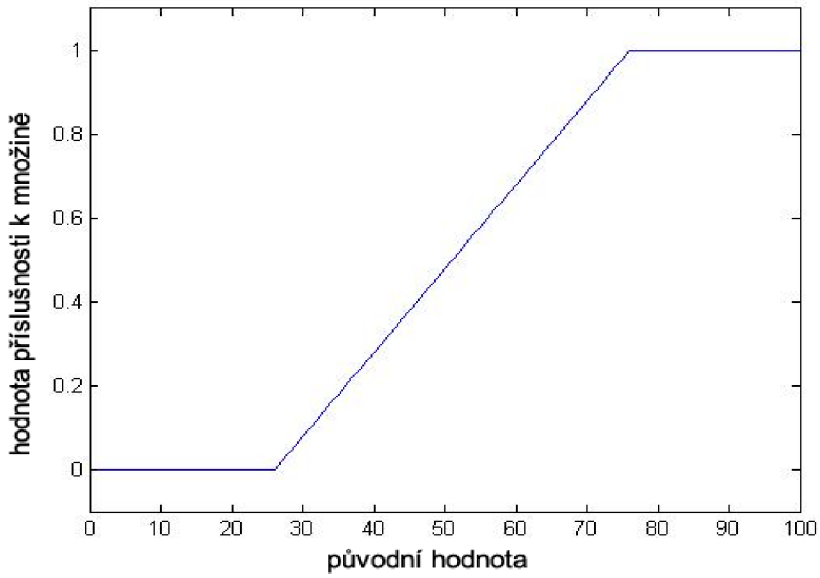


4.4. Zhodnocení výsledků

Z analýzy grafů je zřejmé, že všechny navržené metody rozlišení kódujících a nekódujících úseků sekvencí mají různé výsledky.
U zjišťování obsahu GC bylo zjištěno, že průměrný obsah GC v intronech je 28,18 %, zatímco průměrný obsah v exonech je 50,16 %. Hranice proto byla navržena 39,2 %, což by mohlo být pro rozlišení úseků zcela dostatečné.
U analýzy frekvencí oligonukleotidů výsledky již tak zřejmé nejsou. Práh byl u obou metod navržen s ohledem na dostatečnou diskriminaci a to jako průměr všech hodnot. Pokud bychom u dinukleotidů zvolili například práh jako polovinu maxima, šlo by o hodnotu nižší a z toho důvodu by méně rozdělila kódující a nekódující úseky. V tomto případě by bylo také složité rozhodnout, jestli vzít hodnotu z absolutního maxima nebo maxima z jedné skupiny. Při zvolené hodnotě prahu algoritmus rozdělil 9 z 16 dinukleotidů s nadprahovou a podprahovou hodnotou u intronů či exonů. U trinukleotidů jsme proto postupovali stejně, i když zde by polovina maxima měla hodnotu vyšší. Opět jsme 100 % podělili počtem trinukleotidů a to 64, výsledný práh tak vyšel 1,625. Zde práh rozdělil na první pohled 21 z 64 trinukleotidů.
V analýze se mohou vyskytovat i chyby, které mohou být způsobeny například špatným, či rozdílným sekvenováním DNA, kdy mají sekvence různou délku. V našem souboru se to může projevit hlavně u nekódujících úseků, které mají délku 952 –1386 bází. Tyto rozdíly se pak projeví při zarovnání přidáním mezer a ty pak mají vliv na výsledné frekvence. Tomu se ovšem snažilo předejít spočítáním poměrného výskytu jen z počtu nalezených oligonukleotidů.

4.5. Určení kódujících a nekódujících sekvencí

Pro dosažení lepších výsledků byly navržené metody rozlišení kódujících a nekódujících úseků sekvencí spojeny v jednu funkci, která pomocí fuzzy funkce příslušnosti určí příslušnost ke kódující či nekódující sekvenci pro každou metodu zvlášť. Na jejich základě je nakonec spočítána celková příslušnost.
Fuzzy funkce využívá rovnice pro speciální případ lichoběžníkové funkce - R-funkce (3), kde a je dolní hranice (nulová příslušnost), a b je horní hranice (úplná příslušnost). V našem případě nastavena na 25 a 75 procent. Graf funkce s tímto nastavením je na Obr. 22.



Obr. 22: Lichoběžníková R-funkce příslušnosti použitá pro analýzu

Z předchozí analýzy byly zvoleny jednotlivé dinukleotidy a trinukleotidy, které se v kódujících a nekódujících úsecích nejvíce liší. Dinukleotidů bylo vybráno 8, mezi nimi jsou AA, AT, CC, GC, GG, TA, TT, TG. Čtyři z nich jsou vyšší u exonů a 4 u intronů.
Je spočítán průměr výskytu všech dinukleotidů a tím je stanovena hranice k rozdělení. Pokud je zadaný dinukleotid nad touto hranicí, je exonu či intronu připočten jeden bod.

Tabulka 12: Počty bodů exonů a intronů pro 10 kódujících sekvencí

	seq 1	seq 2	seq 3	seq 4	seq 5	seq 6	seq 7	seq 8	seq 9	seq 10
počet bodů pro exony	**8**	**8**	**8**	**8**	**8**	**8**	**8**	**8**	**7**	**8**
počet bodů pro introny	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

Počty bodů jdou pak na vstup R-funkce příslušnosti, pomocí které je vyjádřeno, jakou příslušnost k jednotlivým množinám mají. Parametry funkce jsou a=2, b=6. Příklad počtu bodů a z nich vypočítanou příslušnost můžeme vidět v následující tabulce.

Tabulka 13: Ukázka vstupních a výstupních hodnot funkce příslušnosti

Počet bodů	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Příslušnost k množině	**0**	**0**	**0,1**	**0,3**	**0,5**	**0,7**	**1**	**1**

Tímto způsobem je vyjádřeno jakou mírou se jedná o kódující či nekódující sekvenci na základě výskytu dinukleotidů.
Obdobným způsobem jsou spočítány trinukleotidy. Zde je vybráno 28 nejvíce se lišících trinukleotidů a opět je určena hranice průměrného výskytu. Jsou spočítány body pro introny

a body pro exony a pomocí funkce příslušnosti spočítána míra příslušnosti k exonům či intronům. Parametry funkce je opět 25 a 75 procent tedy a=7, b=21.

Tabulka 14: Ukázka konkrétních dat bodů intronů u trinukleotidů nekódující sekvence

Počet bodů	23	22	20	22	22	26	25	23	22	23
Příslušnost k intronům	1	1	0,93	1	1	1	1	1	1	1

Posledním parametrem pro rozhodování je obsah GC. Ten jde také na lichoběžníkovou funkci a vyjádřena příslušnost k množině. Jelikož průměr obsahu GC u našich sekvencí byl 39 %, jsou zde jako parametry zvoleny a=28, b=50.
Příslušnosti z jednotlivých metod obsahu dinukleotidů, trinukleotidů a GC jsou zprůměrovány a tím je vyjádřena konečná příslušnost k exonům či intronům, podle toho co chceme zkoumat.

Tabulka 15: Příslušnost kódujících sekvencí k exonům

	seq 1	seq 2	seq 3	seq 4	seq 5	seq 6	seq 7	seq 8	seq 9	seq 10
obsah dinukleotidu	1	1	1	1	1	1	1	1	1	1
obsah trinukleotidu	1	1	1	1	1	1	1	1	1	1
obsah GC	1	1	1	1	1	1	1	1	0,98	0,79
příslušnost	1	1	1	1	1	1	1	1	0,99	0,93

Tabulka 16: Příslušnost kódujících sekvencí k intronům

	seq 1	seq 2	seq 3	seq 4	seq 5	seq 6	seq 7	seq 8	seq 9	seq 10
obsah dinukleotidu	0	0	0	0	0	0	0	0	0	0
obsah trinukleotidu	0	0	0	0	0	0	0	0	0	0
obsah GC	0	0	0	0	0	0	0	0	0,02	0,21
příslušnost	0	0	0	0	0	0	0	0	0,01	0,07

Tabulka 17: Příslušnost nekódujících sekvencí k exonům

	seq 1	seq 2	seq 3	seq 4	seq 5	seq 6	seq 7	seq 8	seq 9	seq 10
obsah dinukleotidu	0	0	0	0	0	0	0	0	0	0
obsah trinukleotidu	0	0	0,07	0	0	0	0	0	0	0
obsah GC	0	0,11	0	0	0	0,26	0,26	0,21	0,05	0
příslušnost	0	0,04	0,02	0	0	0,09	0,09	0,07	0,02	0

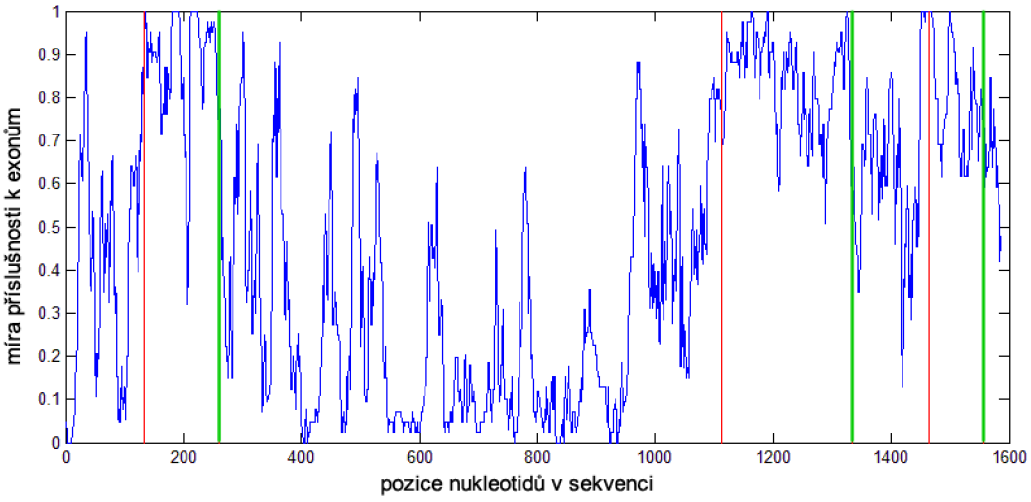
Tabulka 18: Příslušnost nekódujících sekvencí k intronům

	seq 1	seq 2	seq 3	seq 4	seq 5	seq 6	seq 7	seq 8	seq 9	seq 10
obsah dinukleotidu	1	1	1	1	1	1	1	1	1	1
obsah trinukleotidu	1	1	0,93	1	1	1	1	1	1	1
obsah GC	1	0,89	1	1	1	0,74	0,74	0,79	0,95	1
příslušnost	1	0,96	0,98	1	1	0,91	0,91	0,93	0,98	1

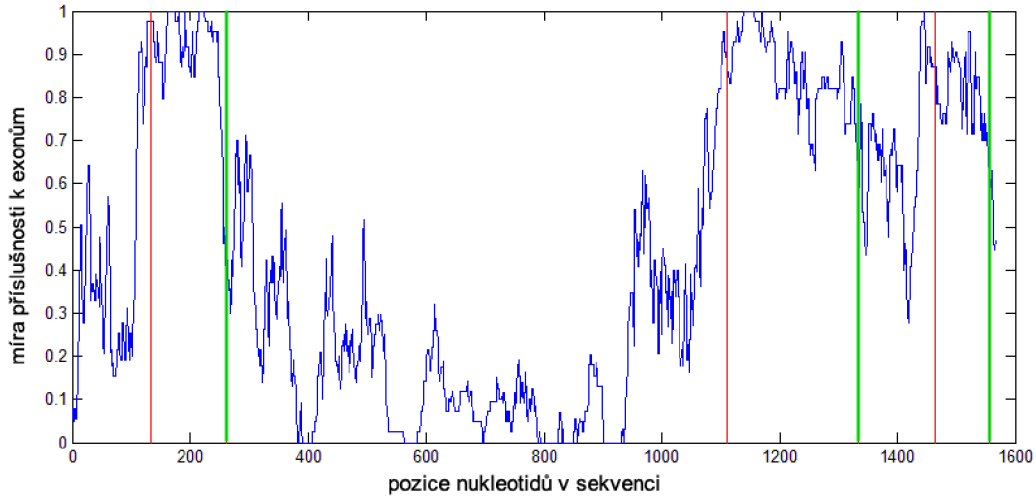
Z výsledků vyplývá, že celkovou příslušnost nad 90 % mají ke své skupině intronů nebo exonů všechny sekvence.

4.6. Hledání kódujících úseků v sekvenci

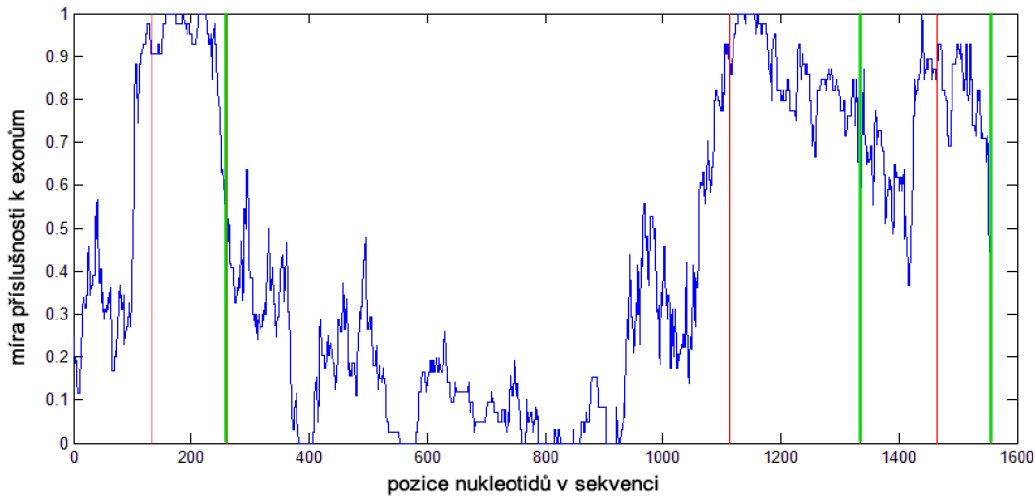
Algoritmus pro určování exonů a intronů lze využít pro zjištění přibližného umístění v celé sekvenci. Při něm je sekvence projížděna oknem zadané délky a pro každé okno spočítáme míru příslušnosti dinukleotidů, trinukleotidů a obsahu GC. Výsledné hodnoty poté závisí na volbě délky okna a také opět na parametrech funkce příslušnosti. Na následujících obrázcích vidíme několik ukázek vynesených v grafu pro jednu sekvenci (*gi|224589802:5246696-5248301 Homo sapiens chromosome 11, GRCh37.p10 Primary Assembly CDS(133..261,1112..1334,1465..1556)*) a rozdílné délky okna. Červené čáry značí začátky kódujících úseků a zelené čáry konce kódujících úseků, které byly získány z databáze EMBL. Modrá čára pak vyjadřuje příslušnost ke kódujícím úsekům (1 úplná příslušnost, 0 žádná). Graf na Obr. 23 je pro délku okna 20, na Obr. 24 pro 40, na Obr. 25 pro 50, na Obr. 26 pro 60 a na Obr. 27 pro délku okna 80.



Obr. 23: Graf příslušnosti k exonům pro délku okna 20



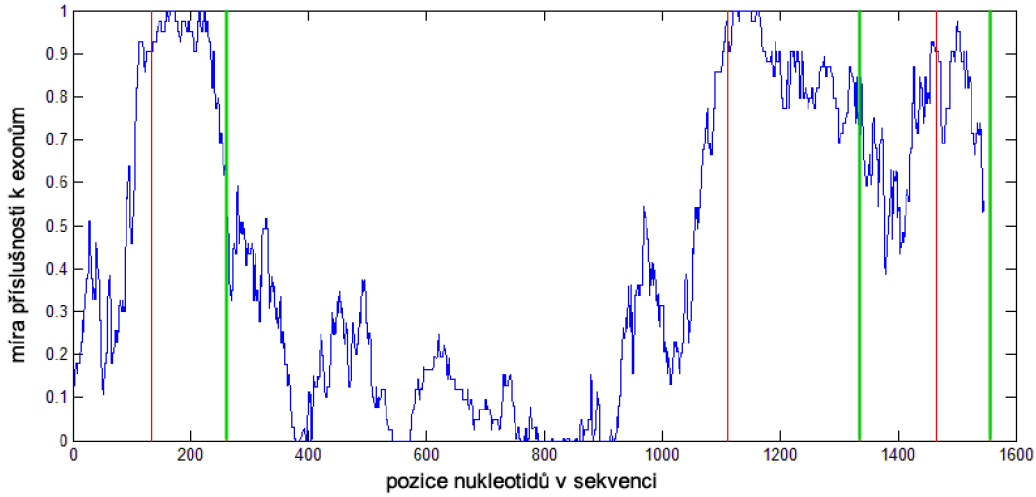
Obr. 24: Graf příslušnosti k exonům pro délku okna 40



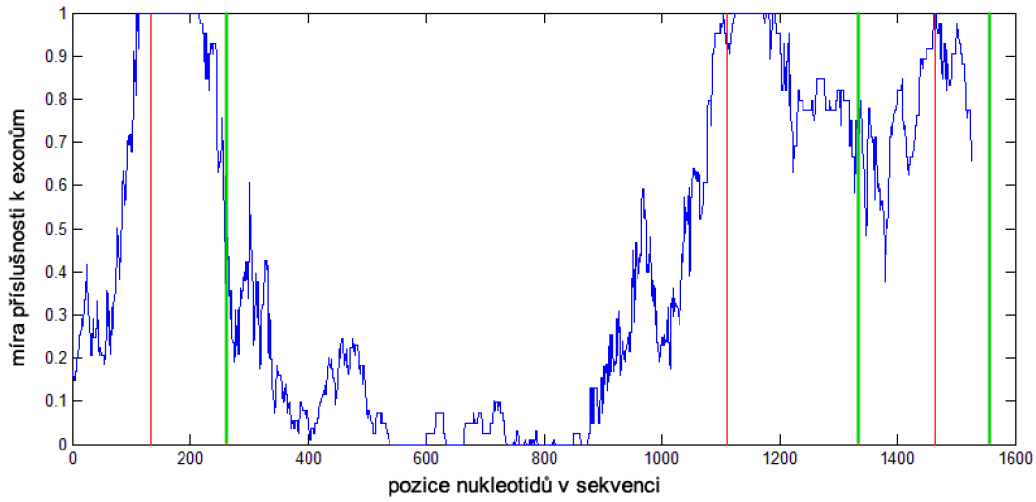
Obr. 25: Graf příslušnosti k exonům pro délku okna 50

Tabulka 19: Citlivost rozlišení exonů a intronů s měnící se délkou okna

délka okna	20	30	40	50	60	70	80	90	100	110	120	130	140	150
Ø exonů	0,844	0,875	0,882	0,890	0,888	0,908	0,909	0,904	0,898	0,899	0,889	0,915	0,915	0,917
Ø intronů	0,342	0,314	0,306	0,307	0,309	0,312	0,312	0,315	0,316	0,324	0,337	0,351	0,360	0,369
citlivost	2,466	2,791	2,885	2,902	2,870	2,914	2,912	2,870	2,840	2,772	2,643	2,603	2,544	2,486

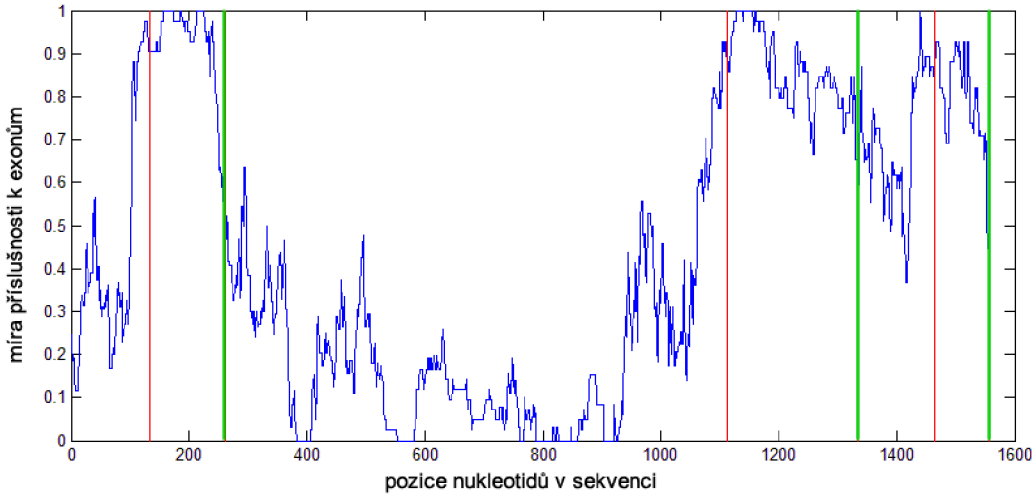


Obr. 26: Graf příslušnosti k exonům pro délku ona 60

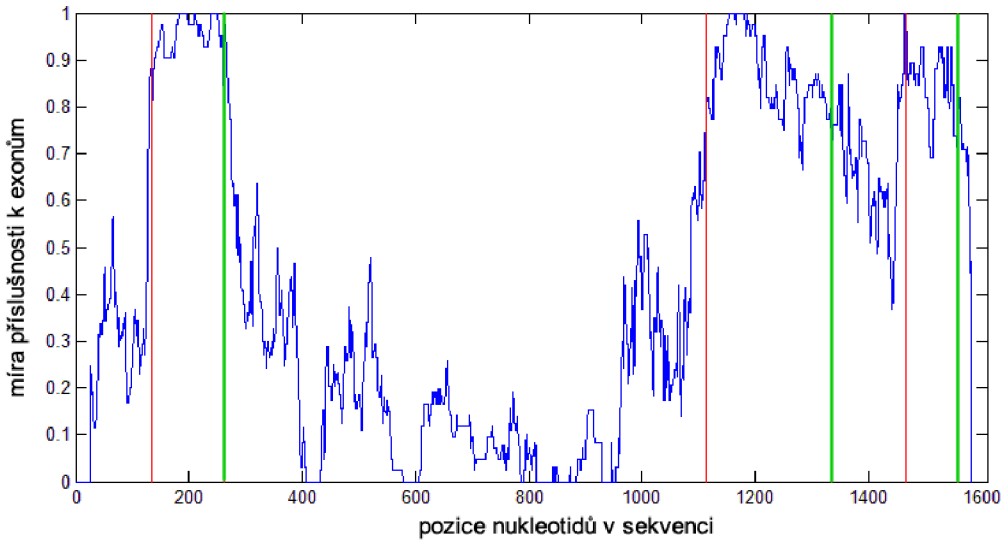


Obr. 27: Graf příslušnosti k exonům pro délku okna 80

Podle grafů to vypadá, že nejlepšího rozlišení je dosaženo u délky okna mezi 40 a 50. U okna délky 20 jsou hodnoty roztříštěné, protože 20 sekvencí je dost krátký úsek pro výpočet frekvencí dinukleotidů, či trinukleotidů. Z Tabulka 19 ovšem vyplývá, že nejlepší citlivosti rozlišení je dosaženo u délky okna 70. Okno délky nad 80 už je méně diskriminativní v ohraničení začátků a konců úseku. Z grafů s větší délkou okna je také více vidět posunutí vůči začátkům a koncům z důvodu náběhu okna. Proto je provedena korekce, kdy je sekvence posunuta o polovinu délky okna doprava. Rozdíl je vidět na Obr. 28 a Obr. 29, kdy první je bez korekce a druhý s korekcí posunutí sekvence.

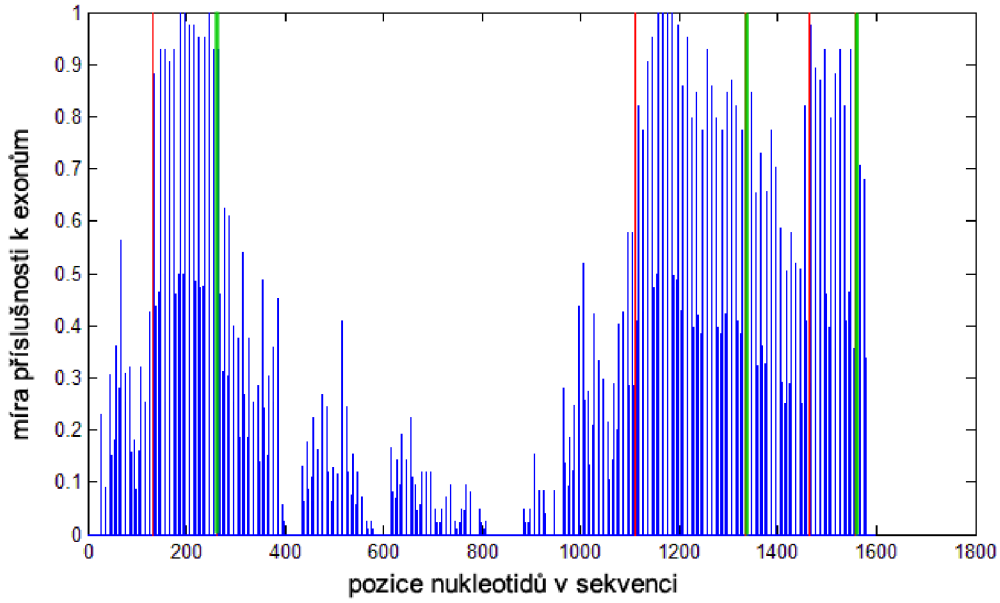


Obr. 28: Graf pro výskyt exonů bez korekce

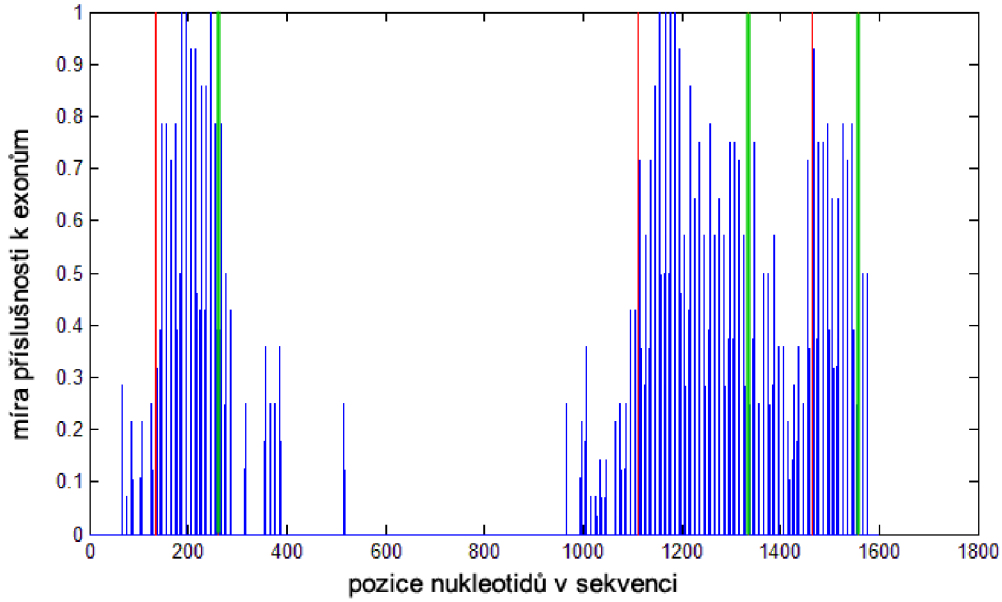


Obr. 29: Graf pro výskyt exonů s korekcí posunutí (pro délku okna 50 posunutí o 25)

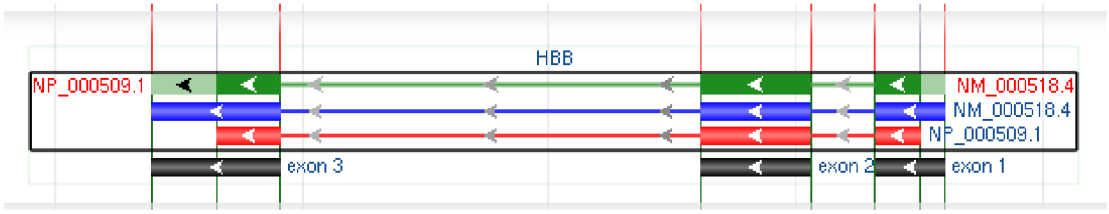
Pro lepší přehlednost je možné okno posunovat ne po 1 nukleotidu ale po 10, kdy získáme méně hodnot. Možné je také volit parametry výpočtu výsledné příslušnosti z 3 příslušností dílčích. Doteď byl vždycky použit průměr těchto 3 hodnot. Na následujících grafech na Obr. 31, Obr. 33 vidíme i hodnoty pro výběr minima a maxima.



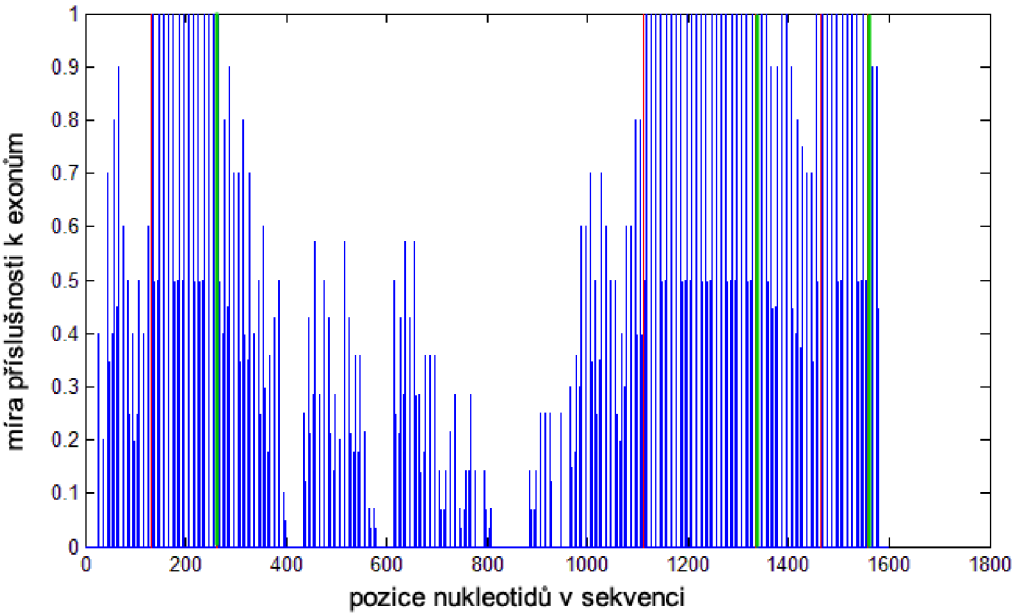
Obr. 30: Průměr příslušností (okno délky 50 se skokem po 10)



Obr. 31: Minimální příslušnost (okno délky 50 se skokem po 10)

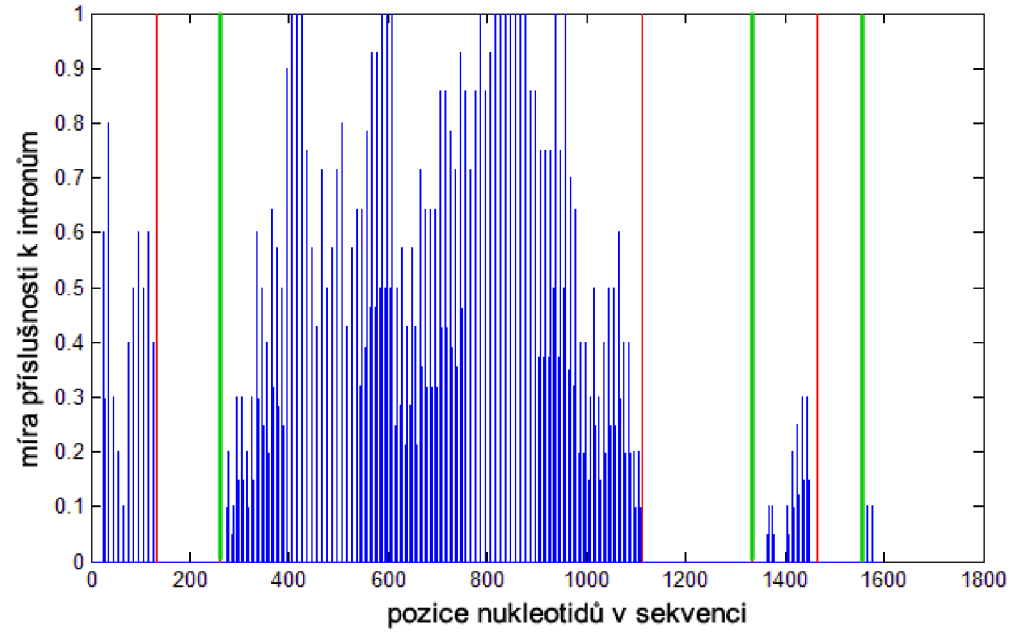


Obr. 32: Grafické vyjádření pozice exonů z databáze NCBI u stejné sekvence



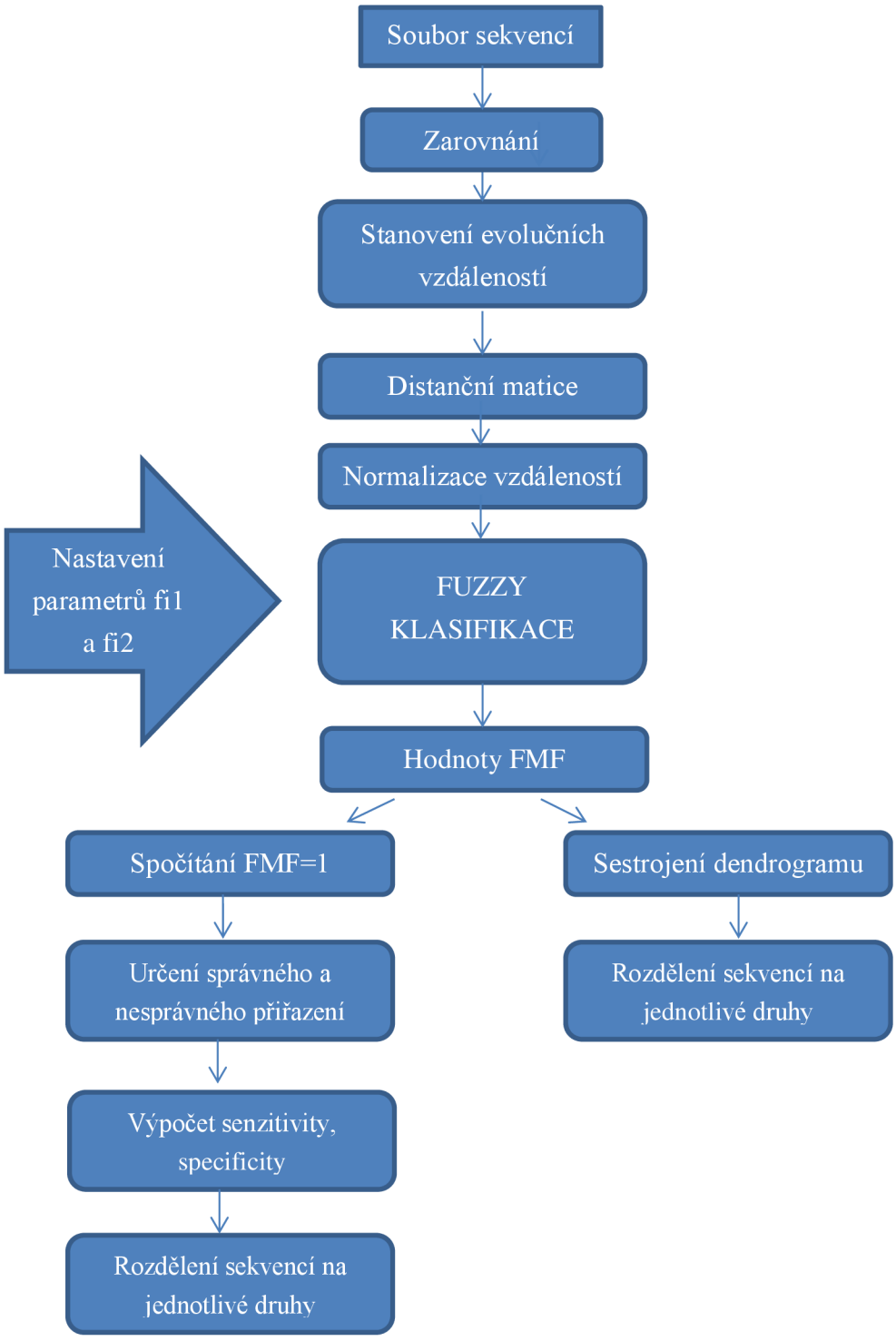
Obr. 33: Maximální příslušnost (okno délky 50 se skokem po 10)

Na posledním Obr. 34 je možné vidět opačný přístup, a to vyhledání intronů tedy nekódujících částí sekvence. Vyhledání intronů se také zdá na první pohled přehlednější. Nejspíše je to způsobeno jejich několikrát větší délkou. Na Obr. 32 je pak grafické vyjádření pozice exonů z databáze NCBI.



Obr. 34: Vyhledání intronů ve stejné sekvenci (prům. přísl., okno délky 50 se skokem po 10)

5. Fuzzy klasifikace sekvencí v DNA barcodingu



Vstupem je soubor sekvencí ve formátu fasta. Ty jsou nejprve zarovnány, je stanovena jejich evoluční vzdálenost, která je vynesena do distanční matice. Hodnoty v této matici jsou normalizovány a jdou na vstup fuzzy funkce, která provede fuzzy klasifikaci a podle nastavených parametrů spočítá hodnoty FMF, které jsou opět v matici stejného rozměru. První možností je sestrojení dendrogramů, z nichž můžeme vyčíst rozdělení jednotlivých druhů. Druhou možností využití matice hodnot FMF je spočítání správných a nesprávných přiřazení, poté senzitivity a specificity rozdělení na jednotlivé druhy.

5.1. Použité sekvence

Sekvence ryb

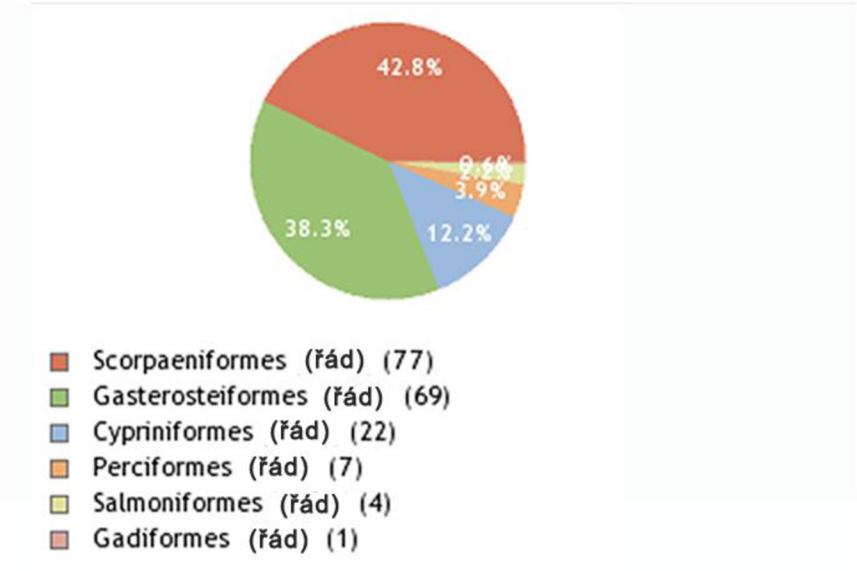
Analýza byla provedena na experimentálním vzorku sekvencí ryb, konkrétně Actinopterygii of Churchil.

Actinopterygii (Paprskoploutví) je velmi početná třída kostnatých ryb, tvořící více než polovinu žijících obratlovců. Moderní systematika obvykle paprskoploutvé dělí na tři nadřády: chrupavčití (Chondrostei), mnohokostnatí (Neopterygii) a kostnatí (Teleostei). Třída má přibližně 42 řádů se 430 čeleděmi a více než 23 000 druhů. Tato skupina ryb tvoří 96 % všech ryb. Celé dvě pětiny všech paprskoploutvých tvoří sladkovodní druhy

Sekvence byly získány z veřejně přístupné databáze barcode sekvencí BOLD [3]. Set obsahuje 180 sekvencí 17 druhů.

Tabulka 20: Zařazení třídy Actinopterygii v říši živočichů

Říše	**živočichové (Animalia)**
Kmen	**strunatci (Chordata)**
Podkmen	**obratlovci (Vertebrata)**
Nadtřída	**ryby (Osteichthyes)**
Třída	**paprskoploutví (Actinopterygii)**



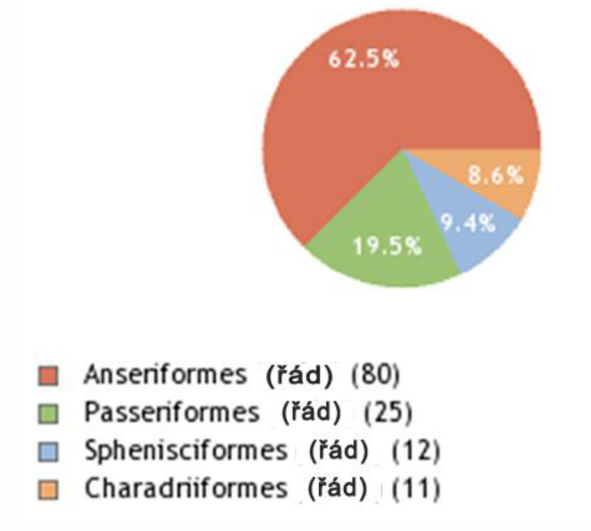
Obr. 35: Rozdělení zástupců do řádů v experimentálním vzorku [3]

Tabulka 21: Počet sekvencí jednotlivých druhů ryb

název druhu	řád	počet sekvencí
Ammodytes hexapterus	Perciformes	3
Catostomus commersonii	Cypriniformes	8
Cottus cognatus	Scorpaeniformes	7
Cottus ricei	Scorpaeniformes	1
Culaea inconstans	Gasterosteiformes	48
Cyclopterus lumpus	Scorpaeniformes	17
Gymnocanthus tricuspis	Scorpaeniformes	11
Lota lota	Gadiformes	1
Lumpenus fabricii	Perciformes	1
Myoxocephalus	Scorpaeniformes	1
Myoxocephalus quadricornis	Scorpaeniformes	18
Myoxocephalus scorpioides	Scorpaeniformes	18
Myoxocephalus scorpius	Scorpaeniformes	4
Pungitius pungitius	Gasterosteiformes	21
Rhinichthys cataractae	Cypriniformes	14
Stichaeus punctatus	Perciformes	3
Thymallus arcticus	Salmoniformes	4

Sekvence ptáků

Sekvence byly získány z veřejně přístupné databáze barcode sekvencí BOLD [3]. Set obsahuje 127 sekvencí 10 druhů.



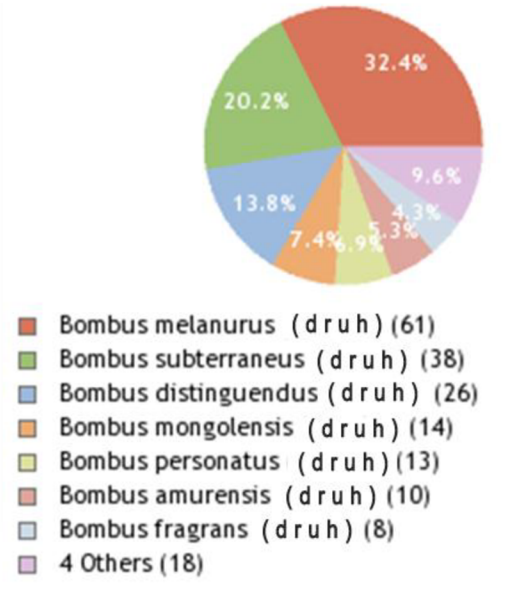
Obr. 36: Rozdělení sekvencí ptáků [3]

Tabulka 22: Počet sekvencí jednotlivých druhů ptáků

druh	počet sekvencí
Branta hutchinsii	80
Eudyptes chrysocome	5
Eudyptes chrysolophus	3
Eudyptes schlegeli	4
Gallinago delicata	3
Gallinago gallinago	5
Gallinago nigripennis	1
Gallinago paraguaiae	1
Progne subis	23
Riparia riparia	2

Sekvence hmyzu

Sekvence byly získány z veřejně přístupné databáze barcode sekvencí BOLD [3]. Set obsahuje 188 sekvencí 11 druhů.



Obr. 37: Rozdělení sekvencí Bumblebees [3]

Tabulka 23: Počet sekvencí jednotlivých druhů hmyzu

druh	počet sekvencí
Bombus amurensis	**8**
Bombus appositus	**4**
Bombus borealis	**5**
Bombus difficillimus	**8**
Bombus distinguendus	**26**
Bombus fedtschenkoi	**1**
Bombus fragrans	**7**
Bombus melanurus	**57**
Bombus mongolensis	**10**
Bombus personatus	**12**
Bombus subterraneus	**38**

5.2. Vyhodnocení správnosti analýzy

Senzitivita testu

Senzitivita testu neboli citlivost testu nabývá hodnot od 0 do 1 (případně 100 %) a vyjadřuje úspěšnost, s níž test zachytí přítomnost sledovaného stavu u daného subjektu.

$$senzitivita = \frac{\text{počet skutečně pozitivních}}{\text{počet skutečně pozitivních+počet falešně negativních}} \tag{13}$$

Specificita testu

Specificita testu vyjadřuje schopnost testu přesně vybrat případy, u nichž zkoumaný znak nenastává.

$$specificita = \frac{\text{počet skutečně negativních}}{\text{počet skutečně negativních + počet falešně pozitivních}} \tag{14}$$

ROC křivka

ROC (Receiver Operating Characteristic) nám vlastně udává hodnocení a optimalizaci klasifikačního systému (testu), který ukazuje vztah mezi specificitou a senzitivitou. Bod na ROC křivce odpovídá hodnotě dělícího kritéria.

Test nebo diagnostická metoda jsou tím užitečnější, čím vyšší je jejich senzitivita a specificita. Nastavením prahových hodnot hledáme na ROC křivce kompromis mezi množstvím falešně pozitivních a falešně negativních výsledků. Ideální ROC křivka stoupá téměř svisle vzhůru (úspěšnost jde ke 100 % a míra chyb zůstává blízko 0 %), teprve pak se zvyšuje míra falešné pozitivity. Naproti tomu, když ROC křivka stoupá po úhlopříčce, znamená to, že každé zlepšení senzitivity je zaplaceno stejně významným zhoršením specificity a test není dobře navržen.

Přibližné zhodnocení kvality testu podle plochy pod křivkou 0,50 až 0,75 = oprávněný, 0,75 až 0,92 = dobrý, 0,92 až 0,97 = velmi dobrý, 0,97 až 1,00 = vynikající. [38]

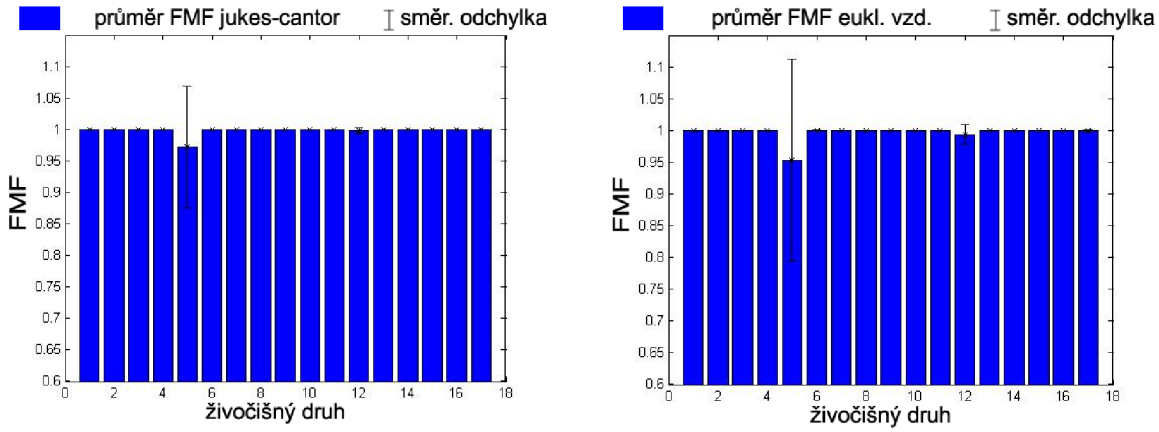
Dále je v práci počítána hodnota ROC max, která vyjadřuje nejlepší společné nastavení senzitivity a specificity, tzn. hodnota, která se nejvíce blíží levému hornímu rohu ROC grafu. Spočítá se jako klasická úhlopříčka obdélníku pomocí pythagorovy věty, kde jsou místo délek stran dosazeny senzitivita a specificita.

5.3. Klasifikace sekvencí

Soubor sekvencí byl po načtení nejprve zarovnán pomocí skórovací matice BLOSUM30. Vzdálenosti mezi jednotlivými sekvencemi byly počítány dvěma způsoby. Prvním je euklidovská vzdálenost z vypočítaných denzit DNA se zadanou délkou okna a druhým je výpočet pomocí běžně používané metody Jukes-Cantor.

5.3.1. Výpočet fuzzy hodnot

Vzdálenosti všech sekvencí, získaných v předchozím kroku, jsou normalizovány od 0 do 1 (poděleny maximální hodnotou) a jdou na vstup funkce příslušnosti, kterou jsou zkorigovány. Výsledkem pro každou vzdálenost je hodnota FMF (fuzzy funkce příslušnosti). Tato funkce je definované v detailu podle rovnice 1. Její parametry Φ1 a Φ2 je třeba odhadnout dle konkrétního souboru dat. V našem případě byl pro ukázku použit 5-tý percentil pro vzdálenost v rámci druhu a 95-tý pro mezidruhovou vzdálenost, který byl vybrán na základě častého používání u DNA barcodingu. Výsledkem je opět matice, která místo vzdáleností obsahuje hodnoty FMF, tedy hodnoty příslušnosti pro jednotlivé sekvence. Pokud se FMF rovná 1, značí to úplnou příslušnost sekvence ke druhé sekvenci, čili zástupce k ostatním zástupcům v rámci druhu. Jelikož vzdálenosti počítáme třemi způsoby, vyjdou i tři distanční matice FMF a to FMF-JC pro výpočet vzdáleností pomocí Jukes-Cantor, FMF-D15 pro výpočet euklidovských vzdáleností z distancí s použitou délkou okna 15 a FMF-D30 pro výpočet euklidovských vzdáleností z distancí s použitou délkou okna 30. Průměr hodnot FMF se směrodatnou odchylkou můžeme vidět na Obr. 38.



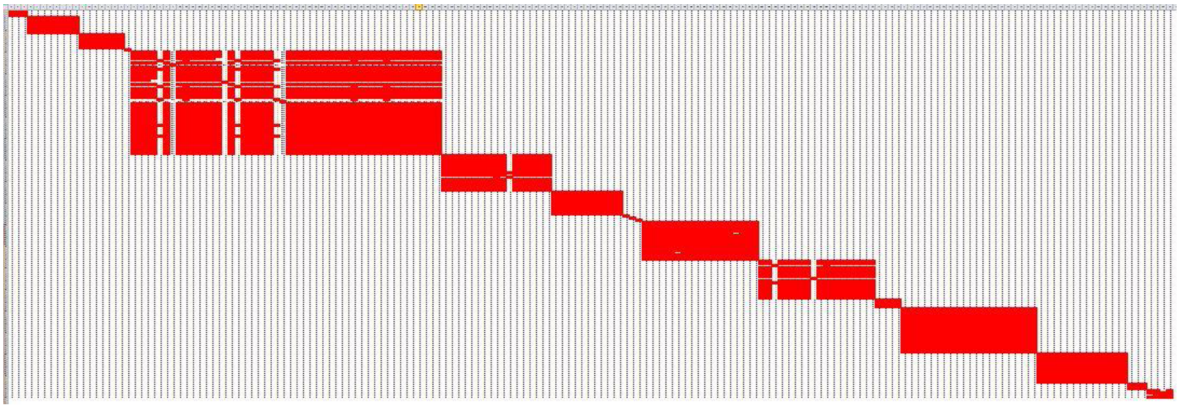
Obr. 38: Průměr FMF se směrodatnou odchylkou pro FMF-JC vlevo a FMF-D30 vpravo

Na obrázcích jsou vyneseny průměrné hodnoty FMF v rámci jednotlivých 17 druhů s vyznačenou směrodatnou odchylkou pro FMF-JC a pro FMF-D30. Tyto hodnoty jsou pro jednotlivé druhy přehledněji zaznamenány i v následující Tabulce.

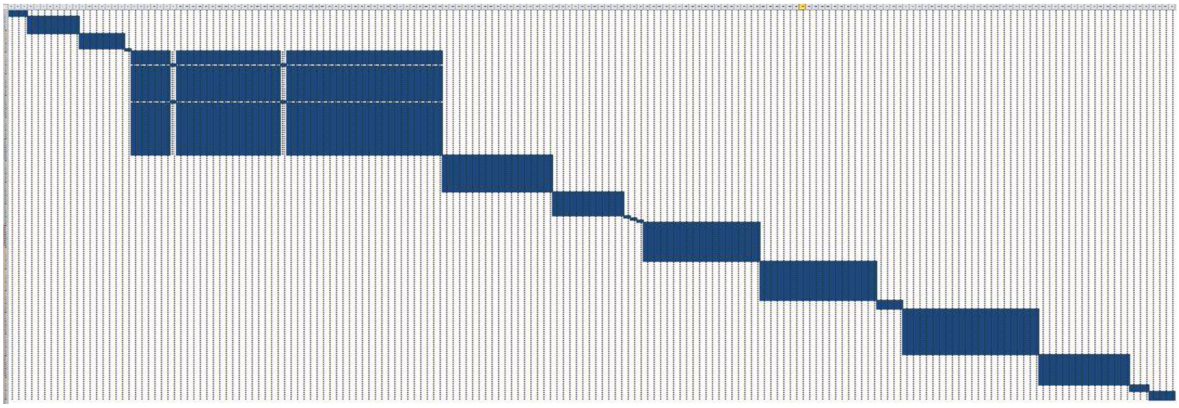
Tabulka 24: Porovnání průměrných hodnot pro JC a Euklidovké vzdálenosti

	FMF-JC					FMF-D30				
název druhu	průměr FMF	směr. odch.	počet sekvencí	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 quadricornis	1	0	18	0	0	1	0	18	0	0
Myoxocephalus scorpioides	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

V tabulce je také zaznamenán počet sekvencí jednotlivých druhů a počty sekvencí, které v rámci druhu neodpovídají úplné příslušnosti (FMF≠1) a ty, které mají příslušnost ke svému druhu menší než 90 %



Obr. 39: Červené vyznačení FMF=1 na celém souboru 180 sekvencí (Euklidovská vzdálenost)



Obr. 40: Modré vyznačení FMF=1 na celém souboru 180 sekvencí (vzdálenost Jukes-Cantor)

Jak vidíme z Obr. 39 a Obr. 40, sekvence byly seřazeny abecedně podle názvů druhů, tzn. zástupci stejného druhu jsou vedle sebe. Matici FMF hodnot prochází cyklus, který spočítá kolikrát je každá sekvence zařazena jako stejná s ostatními (FMF=1). Jsou spočítány správné i špatné přiřazení. V následující tabulce vidíme příklad spočítaných správných a špatných přiřazení FMF-JC pro parametry φ1=0,1 a φ2 =0,9 na části sekvencí:

Tabulka 25: Příklad počítáni úspěšných a neúspěšných přiřazení

číslo sekvence	1	1	1	1	1	1	1	1	1	2	3	4	5
počet sekvencí	18	18	18	18	18	18	18	18	18	11	1	1	1
FMF=1	18	18	17	17	17	17	2	17	17	11	1	1	2
správně přiřazené	18	18	17	17	17	17	1	17	17	11	1	1	1
špatně přiřazené	0	0	0	0	0	0	1	0	0	0	0	0	1

Počty správně a nesprávně přiřazených sekvencí jsou v rámci každého druhu sečteny a vypočítán jejich průměr pro druh jako celek. Z těchto hodnot je spočítána také senzitivita a specificita rozlišení.
V následující tabulce jsou vidět hodnoty spočítané pro nastavení φ1=0,05 a φ2 =0,95 ze vzdáleností FMF-JC.

Z výsledků v Tabulce 23 vidíme, že senzitivita i specificita jsou na dobré úrovni, skoro u všech druhů se blíží jedné. Pro ukázku jsou v další tabulce zobrazeny vypočítané hodnoty pro rozdílné parametry a to φ1=0,4 a φ2 =0,6. V tomto případě už dochází k mnoha přiřazením sekvencí k jiným druhům, a tím je způsobeno zhoršení jak senzitivity, tak specificity.

Tabulka 26: Zobrazení rozložení FMF v rámci všech 180 sekvencí při φ1=0,05 a φ2 =0,95

	počet sek.	FMF=1	správně	špatně	senzitivita	specificita
Ammodytes hexapterus	3,00	3,00	3,00	0,00	1,00	1,00
Catostomus commersonii	8,00	8,00	8,00			
Cottus cognatus	7,00	7,00	7,00			
Cottus ricei	1,00	1,00	1,00			
Culaea inconstans	48,00	42,17	42,17		0,88	
Cyclopterus lumpus	17,00	17,00	17,00		1,00	
Gymnocanthus tricuspis	11,00	11,00	11,00			
Lota lota	1,00	1,00	1,00			
Lumpenus fabricii	1,00	1,00	1,00			
Myoxocephalus	1,00	2,00	1,00	1,00		0,99
Myoxocephalus quadricornis	18,00	18,00	18,00	0,00		1,00
Myoxocephalus scorpioides	18,00	16,17	16,11	0,06	0,90	
Myoxocephalus scorpius	4,00	4,00	4,00	0,00	1,00	
Pungitius pungitius	21,00	21,00	21,00			
Rhinichthys cataractae	14,00	14,00	14,00			
Stichaeus punctatus	3,00	3,00	3,00			
Thymallus arcticus	4,00	4,00	4,00			

Tabulka 27: Zobrazení rozložení FMF v rámci celého setu 180 sekvencí při φ1=0,4 a φ2 =0,6

	počet sek.	FMF=1	správně	špatně	senzitivita	specificita
Ammodytes hexapterus	3,00	3,00	3,00	0,00	1,00	1,00
Catostomus commersonii	8,00	8,00	8,00	0,00		1,00
Cottus cognatus	7,00	8,00	7,00	1,00		0,99
Cottus ricei	1,00	8,00	1,00	7,00		0,96
Culaea inconstans	48,00	44,42	44,42	0,00	0,93	1,00
Cyclopterus lumpus	17,00	17,00	17,00	0,00	1,00	1,00
Gymnocanthus tricuspis	11,00	49,27	11,00	38,27		0,77
Lota lota	1,00	1,00	1,00	0,00		1,00
Lumpenus fabricii	1,00	4,00	1,00	3,00		0,98
Myoxocephalus	1,00	37,00	1,00	36,00		0,80
Myoxocephalus quadricornis	18,00	52,00	18,00	34,00		0,79
Myoxocephalus scorpioides	18,00	50,72	18,00	32,72		0,80
Myoxocephalus scorpius	4,00	50,00	4,00	46,00		0,74
Pungitius pungitius	21,00	21,00	21,00	0,00		1,00
Rhinichthys cataractae	14,00	14,00	14,00	0,00		1,00
Stichaeus punctatus	3,00	4,00	3,00	1,00		0,99
Thymallus arcticus	4,00	4,00	4,00	0,00		1,00

5.3.2. Analýza nastavení parametrů FMF

Abychom dokázali určit optimální nastavení používané fuzzy funkce příslušnosti, je senzitivita a specificita rozlišení spočítána pro různé nastavení φ1 a φ2. Za φ1 jsou dosazovány hodnoty od 0,5 do 0 s krokem 0,05 a zároveň za φ2 hodnoty od 0,5 do 1 se stejným krokem. To můžeme vidět také v následujících tabulkách. Hodnoty senzitivity a specificity jsou vypočteny zprůměrováním senzitivity a specificity pro jednotlivé druhy, tak jak bylo počítáno výše.
V řádku ROC max je pak vyjádřen nejlepší „poměr“ specificity a senzitivity. Poslední sloupec, kdy je nastavení φ1=0 a φ2=1, vyjadřuje hodnoty bez použití FMF.

Tabulka 28: Senzitivita a specificita pro FMF-D15

φ1	0,50	0,45	0,40	0,35	0,30	0,25	0,20	0,15	0,10	0,05	0,00
φ2	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	1,00
FMF=1	56,30	41,72	32,53	28,33	27,02	25,52	22,16	21,92	21,41	18,79	10,27
správně	22,71	22,66	22,19	22,16	22,12	22,12	22,12	21,91	21,40	18,79	10,27
špatné	33,59	19,07	10,34	6,18	4,90	3,40	0,03	0,01	0,01	0,00	0,00
senzitivita	0,998	0,998	0,996	0,995	0,995	0,995	0,995	0,989	0,987	0,921	0,632
specificita	0,827	0,889	0,931	0,957	0,968	0,982	0,999	1,000	1,000	1,000	1,000
ROC max	1,296	1,336	1,363	1,381	1,388	1,398	1,410	1,406	1,405	1,359	1,183

Z tabulky je vidět, že nejlepší hodnoty ROC max a to 1,41 má pro FMF-D15 nastavení φ1=0,2 a φ2=0,8

Tabulka 29: Senzitivita a specificita pro FMF-D30

φ1	0,5	0,45	0,4	0,35	0,3	0,25	0,2	0,15	0,1	0,05	0
φ2	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95	1
FMF=1	137,03	97,68	43,14	31,83	27,73	26,04	24,66	21,87	21,30	18,10	10,27
správně	22,66	22,21	22,16	22,12	22,12	22,12	22,11	21,86	21,29	18,10	10,27
špatné	114,38	75,47	20,99	9,71	5,61	3,92	2,54	0,01	0,01	0,00	0,00
senzitivita	0,998	0,996	0,995	0,995	0,995	0,995	0,995	0,989	0,979	0,907	0,632
specificita	0,346	0,603	0,873	0,939	0,960	0,977	0,990	1,000	1,000	1,000	1,000
ROC max	1,056	1,164	1,324	1,368	1,383	1,395	1,404	1,406	1,399	1,350	1,183

Pro FMF-D30 je nejlepší hodnota ROC max 1,406 pro nastavení φ1=0,15 a φ2=0,85.

Tabulka 30: Senzitivita a specificita pro FMF-JC

φ1	0,50	0,45	0,40	0,35	0,30	0,25	0,20	0,15	0,10	0,05	0,00
φ2	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	1,00
FMF=1	171,28	145,06	96,19	61,58	44,27	33,14	29,41	25,90	24,53	19,77	10,27
správně	23,13	22,69	22,62	22,19	22,12	22,12	22,12	22,12	21,84	19,76	10,27
špatné	148,14	122,37	73,57	39,39	22,14	11,02	7,29	3,78	2,69	0,01	0,00
senzitivita	1,000	0,998	0,998	0,996	0,995	0,995	0,995	0,995	0,989	0,961	0,632
specificita	0,069	0,225	0,500	0,780	0,879	0,927	0,954	0,983	0,990	1,000	1,000
ROC max	1,002	1,023	1,116	1,265	1,328	1,360	1,378	1,399	1,399	1,387	1,183

Pro FMF-JC je nejlepší hodnota ROC max 1,399 stejná při dvou nastaveních φ1=0,15 a φ2=0,85 a při nastavení φ1=0,1 a φ2=0,9.

Analýza sekvencí ptáků

Výpočet senzitivity a specificity pro různé nastavení φ1 a φ2 u 127 sekvencí ptáků, rozdělených do 10 druhů. Jejich rozdělení je vidět z grafu na Obr. 36.

Tabulka 31: Senzitivita a specificita pro FMF-D15

φ1	0,50	0,45	0,40	0,35	0,30	0,25	0,20	0,15	0,10	0,05	0,00
φ2	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	1,00
FMF=1	74,83	64,37	56,69	56,26	55,88	55,17	54,04	48,31	33,00	20,53	7,16
správně	55,27	55,27	55,20	55,13	54,98	54,35	53,39	47,93	32,83	20,50	7,16
špatné	19,56	9,10	1,48	1,13	0,90	0,82	0,65	0,38	0,17	0,03	0,00
senzitivita	1,000	1,000	0,954	0,944	0,941	0,849	0,821	0,796	0,707	0,555	0,399
specificita	0,838	0,902	0,952	0,960	0,967	0,969	0,973	0,985	0,995	0,999	1,000
ROC max	1,305	1,347	1,348	1,346	1,349	1,288	1,273	1,267	1,220	1,143	1,077

Pro FMF-D15 vyšel nejlepší ROC max pro φ1=0,3 a φ2=0,7.

Tabulka 32: Senzitivita a specificita pro FMF-D30

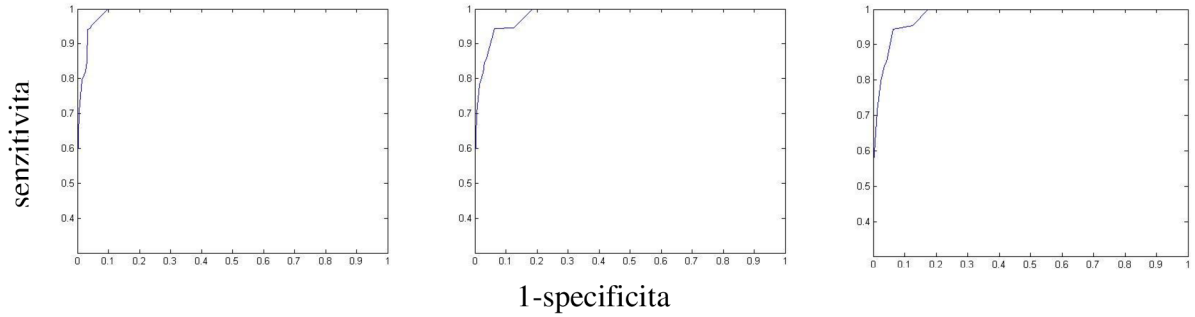
φ1	0,50	0,45	0,40	0,35	0,30	0,25	0,20	0,15	0,10	0,05	0,00
φ2	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	1,00
FMF=1	86,10	79,27	69,82	58,92	55,98	54,73	52,72	45,76	31,52	20,34	7,16
správně	55,27	55,27	55,19	55,08	54,92	53,91	52,04	45,39	31,36	20,31	7,16
špatné	30,83	24,00	14,63	3,84	1,06	0,82	0,68	0,36	0,16	0,03	0,00
senzitivita	1,000	1,000	0,946	0,943	0,860	0,847	0,816	0,784	0,701	0,554	0,399
specificita	0,745	0,814	0,874	0,936	0,963	0,969	0,973	0,985	0,995	0,999	1,000
ROC max	1,247	1,289	1,288	1,329	1,291	1,287	1,270	1,259	1,217	1,143	1,077

Pro FMF-D30 vyšel nejlepší ROC max pro φ1=0,35 a φ2=0,65.

Tabulka 33: Senzitivita a specificita pro FMF-JC

φ1	0,50	0,45	0,40	0,35	0,30	0,25	0,20	0,15	0,10	0,05	0,00
φ2	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	1,00
FMF=1	84,80	78,81	70,29	60,48	55,72	54,97	52,97	47,03	32,65	20,72	7,16
správně	55,27	55,27	55,20	55,09	54,53	53,99	52,12	46,32	32,29	20,67	7,16
špatné	29,53	23,54	15,09	5,39	1,20	0,98	0,85	0,71	0,36	0,05	0,00
senzitivita	1,000	1,000	0,954	0,944	0,852	0,841	0,822	0,797	0,712	0,566	0,399
specificita	0,766	0,826	0,875	0,938	0,959	0,965	0,968	0,975	0,988	0,999	1,000
ROC max	1,259	1,297	1,294	1,330	1,283	1,280	1,270	1,259	1,218	1,148	1,077

U FMF-JC vyšly hodnoty podobné jako pro FMF-D30.



Obr. 41: ROC křivky pro FMF-D15, FMF-D30 a FMF-JC

Na Obr. 41 vidíme ROC křivky pro jednotlivé případy výpočtů.

Tabulka 34: Seznam druhů a hodnoty pro jednotlivé druhy pro nejlepší senzitivitu a specificitu (φ1=0,35 a φ2=0,75)

	počet sek.	FMF=1	správně	Špatně	senzitivita	specificita
Branta hutchinsii	80,00	80,00	80,00	0,00	1,000	1,000
Eudyptes chrysocome	5,00	8,00	3,40	4,60	0,680	0,962
Eudyptes chrysolophus	3,00	7,33	2,33	5,00	0,778	0,960
Eudyptes schlegeli	4,00	10,00	4,00	6,00	1,000	0,951
Gallinago delicata	3,00	10,00	3,00	7,00	1,000	0,944
Gallinago gallinago	5,00	9,80	5,00	4,80	1,000	0,961
Gallinago nigripennis	1,00	10,00	1,00	9,00	1,000	0,929
Gallinago paraguaiae	1,00	9,00	1,00	8,00	1,000	0,937
Progne subis	23,00	23,09	22,65	0,43	0,985	0,996
Riparia riparia	2,00	7,00	2,00	5,00	1,000	0,960

Analýza sekvencí hmyzu

Výpočet senzitivity a specificity pro různé nastavení φ1 a φ2 u souboru sekvencí Bumblebees obsahujícím 188 sekvencí 11 druhů znázorněných na Obr. 37

Tabulka 35: Senzitivita a specificita pro FMF-JC

fi1	0,5	0,45	0,4	0,35	0,3	0,25	0,2	0,15	0,1	0,05	0
fi2	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95	1
FMF=1	151,17	137,98	128,17	115,34	105,22	74,51	45,41	30,45	22,24	15,68	7,33
správně	32,07	31,01	30,86	30,05	29,58	27,17	24,66	24,20	22,00	15,68	7,33
špatné	119,10	106,97	97,31	85,30	75,64	47,34	20,75	6,25	0,24	0,00	0,00
senzitivita	0,990	0,976	0,975	0,962	0,934	0,902	0,775	0,745	0,692	0,648	0,435
specificita	0,230	0,314	0,380	0,470	0,534	0,717	0,849	0,931	0,997	1,000	1,000
ROC max	1,017	1,025	1,047	1,070	1,076	1,153	1,150	1,192	1,213	1,192	1,090

Zde vyšel nejlepší ROC max pro nastavení φ1=0,1 a φ2=0,9 a v následující tabulce jsou pro toto nastavení vyneseny hodnoty senzitivity a specificity pro jednotlivé druhy.

Tabulka 36: Seznam druhů a hodnoty pro jednotlivé druhy pro nejlepší senzitivitu a specificitu (φ1=0,1 a φ2=0,9)

	počet sek.	FMF=1	správně	špatně	senzitivita	specificita
Bombus amurensis	8,00	2,25	2,25	0,00	0,28	1,00
Bombus appositus	4,00	9,00	4,00	5,00	1,00	0,97
Bombus borealis	5,00	5,00	5,00	0,00	1,00	1,00
Bombus difficillimus	8,00	5,13	5,00	0,13	0,63	1,00
Bombus distinguendus	26,00	15,85	15,08	0,77	0,58	0,99
Bombus fedtschenkoi	1,00	1,00	1,00	0,00	1,00	1,00
Bombus fragrans	7,00	4,14	4,14	0,00	0,59	1,00
Bombus melanurus	57,00	37,18	37,18	0,00	0,65	1,00
Bombus mongolensis	10,00	6,70	6,60	0,10	0,66	1,00
Bombus personatus	12,00	5,50	5,50	0,00	0,46	1,00
Bombus subterraneus	38,00	28,95	28,95	0,00	0,76	1,00

Společná analýza sekvencí ryb a hmyzu

Výpočet senzitivity a specificity pro různé nastavení φ1 a φ2 byl vyzkoušen i na spojení 2 setů odlišných sekvencí. V tomto případě výše popisovanými sekvencemi ryb a hmyzu. Tento souhrnný set obsahoval 368 sekvencí 28 druhů. ROC max vyšlo nejvyšší u nastavení parametrů φ1=0,1 a φ2=0,9. V tabulce 37 vidíme, že u různých druhů vycházela senzitivita a specificita různě. Nelze pozorovat nějaký trend růstu hodnot například s počtem sekvencí.

Tabulka 37: Senzitivita a specificita pro FMF-JC

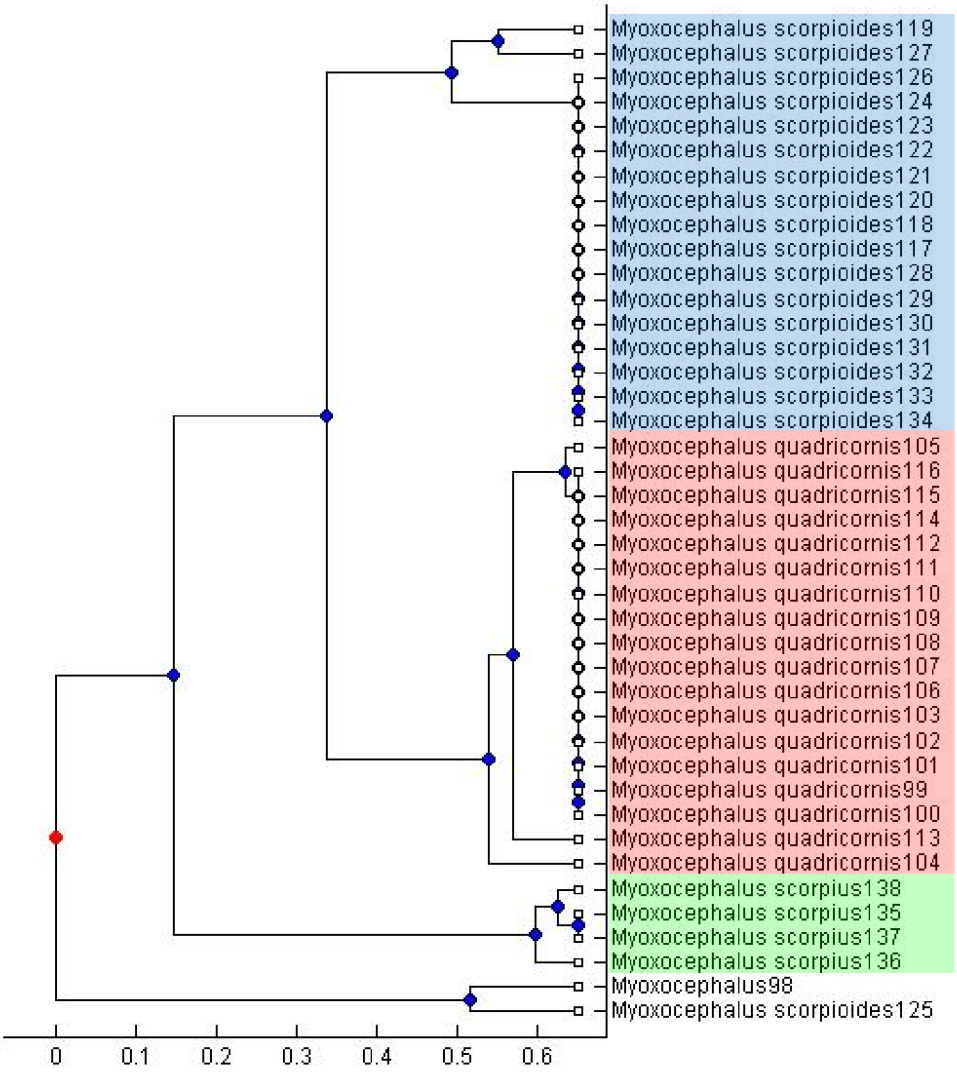
fi1	0,5	0,45	0,4	0,35	0,3	0,25	0,2	0,15	0,1	0,05	0
fi2	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95	1
FMF=1	176,55	176,36	173,60	155,18	123,61	111,42	103,20	92,43	71,86	29,79	8,81
správně	28,07	28,06	27,58	25,94	23,43	21,69	19,87	17,94	15,94	11,66	4,78
špatné	148,48	148,30	146,01	129,24	100,19	89,73	83,33	74,49	55,92	18,12	4,03
senzitivita	0,987	0,987	0,982	0,962	0,916	0,899	0,892	0,875	0,825	0,710	0,477
specificita	0,524	0,524	0,529	0,574	0,656	0,686	0,706	0,745	0,811	0,931	0,980
ROC max	1,117	1,117	1,116	1,120	1,127	1,131	1,138	1,150	1,156	1,171	1,090

Tabulka 38: Seznam druhů a hodnoty pro jednotlivé druhy pro nejlepší senzitivitu a specificitu (φ1=0,1 a φ2=0,9)

	počet sek.	FMF=1	správně	špatně	senzitivita	specificita
Ammodytes hexapterus	3,00	3,00	3,00	0,00	1,00	1,00
Catostomus commersonii	8,00	7,00	3,75	3,25	0,47	0,99
Cottus cognatus	7,00	51,86	7,00	44,86	1,00	0,87
Cottus ricei	1,00	33,00	1,00	32,00	1,00	0,91
Culaea inconstans	48,00	35,13	16,50	18,63	0,34	0,94
Cyclopterus lumpus	17,00	51,47	11,71	39,76	0,69	0,88
Gymnocanthus tricuspis	11,00	51,00	7,73	43,27	0,70	0,87
Lota lota	1,00	62,00	1,00	61,00	1,00	0,83
Lumpenus fabricii	1,00	62,00	1,00	61,00	1,00	0,83
Myoxocephalus	1,00	2,00	1,00	1,00	1,00	1,00
Myoxocephalus quadricornis	18,00	53,72	14,67	39,06	0,81	0,88
Myoxocephalus scorpioides	18,00	38,17	7,33	30,83	0,41	0,91
Myoxocephalus scorpius	4,00	6,50	2,00	4,50	0,50	0,99
Pungitius pungitius	21,00	34,24	14,05	20,19	0,67	0,94
Rhinichthys cataractae	14,00	34,57	9,00	25,57	0,64	0,93
Stichaeus punctatus	3,00	30,67	1,67	29,00	0,56	0,92
Thymallus arcticus	4,00	33,75	2,50	31,25	0,63	0,91
Bombus amurensis	8,00	7,88	6,25	1,63	0,78	1,00
Bombus appositus	4,00	7,00	4,00	3,00	1,00	0,99
Bombus borealis	5,00	21,00	1,80	19,20	0,36	0,95
Bombus difficillimus	8,00	38,63	6,00	32,63	0,75	0,91
Bombus distinguendus	26,00	41,08	21,92	19,15	0,84	0,94
Bombus fedtschenkoi	1,00	45,00	1,00	44,00	1,00	0,88
Bombus fragrans	7,00	44,57	7,00	37,57	1,00	0,89
Bombus melanurus	57,00	16,98	13,39	3,60	0,23	0,99
Bombus mongolensis	10,00	14,70	8,20	6,50	0,82	0,98
Bombus personatus	12,00	13,67	4,00	9,67	0,33	0,97
Bombus subterraneus	38,00	15,16	13,37	1,79	0,35	0,99

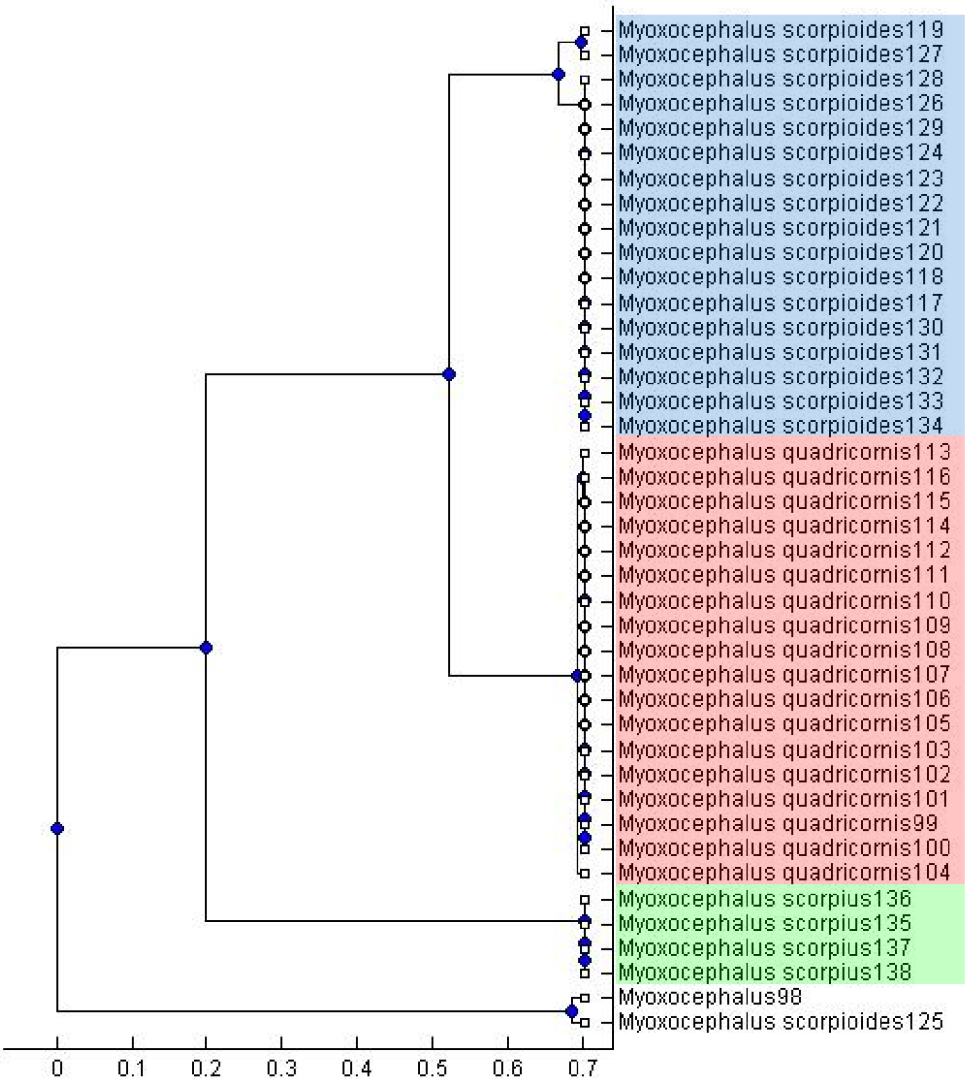
5.3.3. Sestrojení dendrogramů

Z hodnot FMF je možné sestrojit dendrogramy. Na následujících obrázcích vidíme dendrogramy sestavené pomocí UPGMA [25] pro vzdálenosti sekvence. Pro lepší přehlednost jen pro 3 druhy a to *Myoxocephalus quadricornis, Myoxocephalus scorpioides, a Myoxocephalus scorpius*. V prvním případě sestrojené ze vzdáleností bez použití FUZZY funkce v druhém případě s jejím použitím s nastavenými hodnotami φ1=0,1 a φ2=0,9.



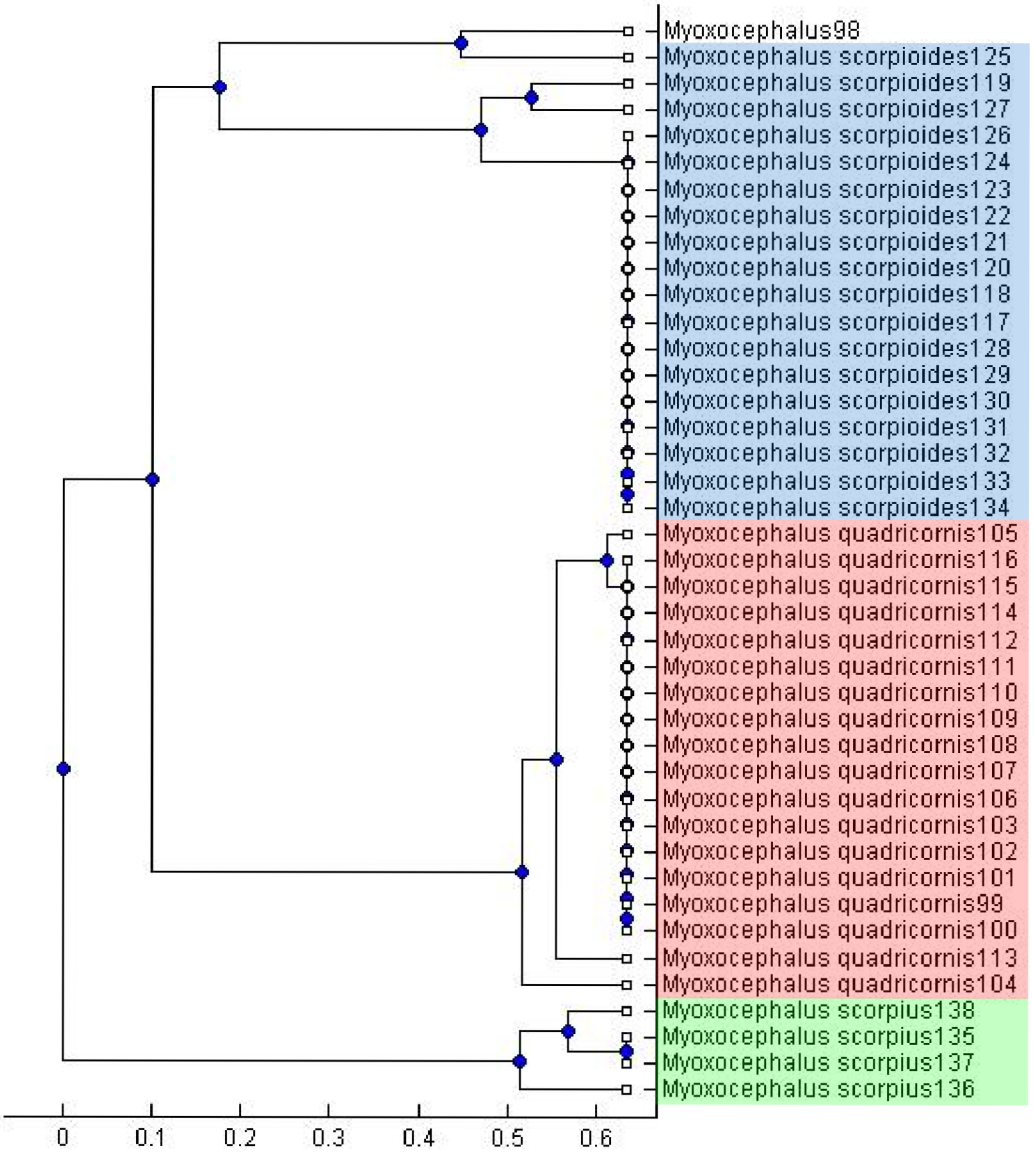
Obr. 42: Dendrogram metodou JC bez použití FMF

Na Obr. 42 je dendrogram sestrojený ze vzdáleností JC a na následujícím Obr. 43 sestrojený z FMF-JC.

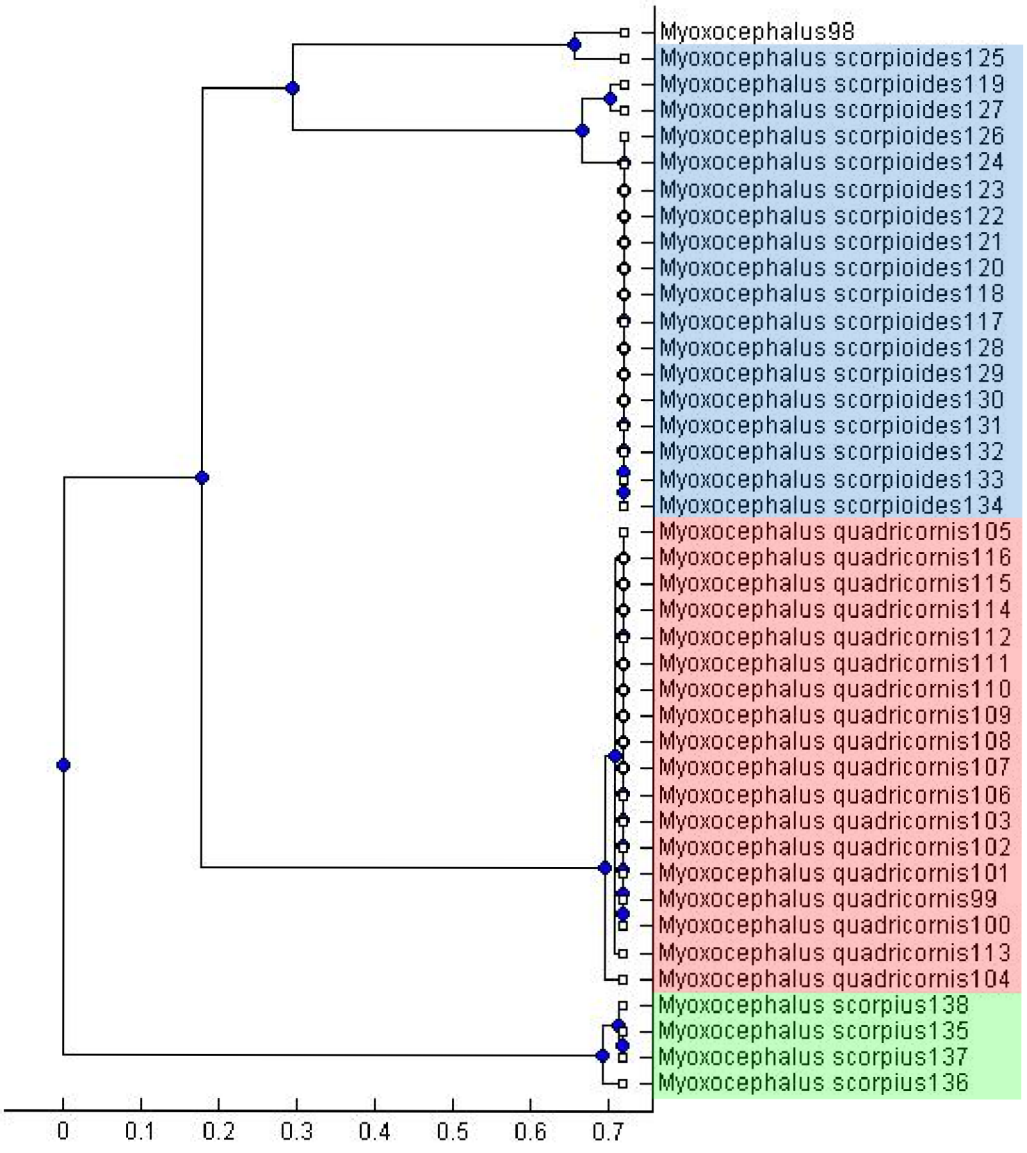


Obr. 43: Dendrogram metodou JC s použitím FMF

Na dalších dvou obrázcích Obr. 44, Obr. 45 je to samé, v tomto případě pro euklidovské vzdálenosti z denzit s délkou okna 30. V prvním případě bez použití FMF a v druhém s použitím FMF(FMF-D30).



Obr. 44: Dendrogram metodou Euklidovské vzdálenosti bez použití FMF



Obr. 45: Dendrogram metodou Euklidovské vzdálenosti s použitím FMF (FMF-D30).

Jak je na první pohled zřejmé, u dendrogramu sestrojeného z upravených hodnot dochází k jeho zjednodušení. Vzdálenosti mezi jednotlivými zástupci v rámci druhu jsou menší, naopak se prodlouží vzdálenosti mezi druhy. Takovýto dendrogram je pak více diskriminativní a lépe interpretovatelný.

Závěr

Úkolem této diplomové práce bylo shrnutí a prohloubení možností využití fuzzy logiky ke klasifikaci DNA sekvencí. K tomuto účelu je v práci nejprve sestaven přehled základních fuzzy klasifikačních kritérií a to GC obsah, frekvence oligonukleotidů, sekvenční motivy a evoluční vzdálenost sekvencí. Teoreticky je představena technika DNA barcodingu sloužící ke klasifikaci a identifikaci organismů, kde fuzzy klasifikace může velkým dílem přispět k zařazení organismů, jejichž příslušnost je kvůli nejasným hranicím mezi druhy problematická. Podrobně jsou zde rozebrány různé skupiny testovacích dat získaných z veřejných databází, na jejichž specifikace je konkrétní typ fuzzy klasifikace zacílen. Z přehledu informací je zřejmé, že jádro samotné metody, a to fuzzy logika, je pojem poměrně mladý, ale do budoucna by mohl mít stále širší uplatnění.

Dalším úkolem, teď už praktickým, bylo formulování klasifikačního algoritmu pro posouzení podobnosti DNA sekvencí a jeho otestování na souboru DNA sekvencí, získaných z veřejných databází.
Pro analýzu bylo zvoleno 10 sekvencí, nejvíce zastoupených homo sapiens a mus musculus, kódujících hemoglobin, konkrétně jeho část beta. Tyto sekvence byly rozděleny na introny a exony a u nich poté provedeny výpočty podílu GC, frekvencí dinukleotidů, a frekvencí trinukleotidů. Na jejich základě bylo navrhnuto pro každou metodu zvlášť kritérium odlišení kódujích a nekódujících úseků, a to práh výskytu. U obsahu GC je rozlišení díky prahu jednoznačné, ale detekce kódujících úseků pouze na základě jednoho triviálního parametru se ukázala nedostatečná. Nastavení prahu pro dinukleotidy je však specifické pro každý dinukleotid a to stejné se po analýze ukázalo i pro frekvence trinukleotidů, kde hodnoty pro některý trinukleotid leží obě dvě nad prahem a přitom se řádově liší. Nemožnost nastavení jednotných klasifikačních kritérií tak poukazuje právě na vhodnost fuzzy klasifikace pomocí vhodně navržené funkce příslušnosti.
Pro dosažení lepších výsledků byly tyto navržené metody pro rozlišení kódujících a nekódujících úseků sekvencí spojeny v jednu funkci, která pomocí fuzzy funkce příslušnosti určí příslušnost ke kódující či nekódující sekvenci pro každou metodu zvlášť. Na jejich základě byla nakonec spočítána celková hodnota příslušnosti, buď ke kódujícím, nebo nekódujícím úsekům. Tato příslušnost se téměř u všech sekvencí blíží 1. Celkovou příslušnost nad 90 % mají ke své skupině intronů nebo exonů všechny sekvence.
Této metody bylo také využito pro algoritmus, který vyhledává začátky a konce kódujících či nekódujících částí sekvence. Ta je zde projížděna oknem o zadané délce, v našem případě bylo testováno okno délky 20 – 150, kdy nejlepších výsledků bylo dosahováno u okna velikosti 70.

V tomto okně jsou počítány míry příslušnosti daného úseku ke kódující nebo nekódující sekvenci. Určení není zcela přesné, což vyplývá už z fuzzy podstaty metody, ale přibližné rozlišení je z grafů zřetelné.

Další zde rozpracovanou metodou je využití fuzzy v DNA barcodingu. Kombinace fuzzy logiky společně s DNA barcodingem je biologicky přirozenější klasifikací sekvencí využívající neostrých hranic mezi jednotlivými druhy. Spojení tohoto přístupu využití fuzzy klasifikace v DNA barcodingu bylo demonstrováno na experimentálním vzorku sekvencí ryb, ptáků a hmyzu. První obsahoval 180 sekvencí od 17 druhů, druhý 127 sekvencí od 10 druhů a poslední zmiňovaný 188 sekvencí od 11 druhů organismů. Byla použita fuzzy funkce, jejíž parametry je třeba nastavit dle konkrétního souboru dat. Evoluční vzdálenosti pro DNA barcoding byly stanoveny na základě euklidovských vzdáleností průběhů nukleotidových denzit a pomocí Jukes-Cantorova evolučního modelu pro nukleotidy. Díky využití fuzzy funkce došlo při vhodně nastavených parametrech ke zlepšení senzitivity klasifikace u souboru hmyzu o 20 %, u ryb o 40 % a u souboru ptáků až o 50 %.
Hodnota ROC max, která určuje nejlepší poměr mezi senzitivitou se specificitou rozlišení příslušnosti sekvence k určitému druhu, se zvýšila z 1,183 na 1,399 u sekvencí ryb, z 1,077 na 1,330 u ptáků, z 1,090 na 1,213 u hmyzu a z 1,090 na 1,171 u spojení sekvencí ryb a hmyzu. Toto zlepšení však záleží hlavně na optimálně zvolených parametrech fuzzy funkce, které jsou získány právě vyhodnocením různých nastavení pomocí ROC max.
Nakonec byly z vypočítaných příslušností sestrojeny dendrogramy. Rozdíl před a po použití FMF ve vykreslených dendrogramech je patrný na první pohled, a to ve zjednodušení struktury a jednodušší interpretaci.

Seznam použité literatury

[1] ALONSO, S, K. Fuzzy operators. eMathTeacher: Mamdani's Fuzzy Inference Method. Retrieved from http://www.dma.fi.upm.es/java/fuzzy/fuzzyinf/funpert_en.htm

[2] ANZALDI LJ, MUÑOZ-FERNÁNDEZ D, ERILL I. (2012). "BioWord: a sequence manipulation suite for Microsoft Word". BMC Bioinformatics 13 (124). DOI:10.1186/1471-2105-13-124.PMID 22676326

[3] BOLD. The Barcode of Life Data System, www.barcodinglife.org, 2013.

[4] D'HAESELEER P, What are DNA sequence motifs?, NATURE BIOTECHNOLOGY, VOLUME 24, NUMBER 4, Nature Publishing Group, 2006

[5] DAVID P. CLARK, Molecular biology: Understanding the genetic revolution. Elsevier Academic Press (2005), ISBN 0-12-175551-7

[6] FATIMA CVRČKOVÁ, Úvod do praktické bioinformatiky, Academia, ISBN 80-200-1360-1, 2006

[7] GARCIA, Fernando, Francisco J. LOPEZ, Carlos CANO a Armando BLANCO. Study of fuzzy resemblance measures for DNA motifs. IEEE International Conference on Fuzzy Systems. Korea, 2009, s. 1175-1180

[8] HONG, T., LEEB, C. (1996). Induction of fuzzy rules and membership functions from training examples, *84*(95)

[9] Jiří Hřebíček, Jan Žižka, Vědecké výpočty v biologii a biomedicíně, Masarykova univerzita, 2007

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

[11] KALUŽA, R. Fuzzy logika, 2006, [online], [cit. 2012-11-01]. Dostupné z WWW: http://radovan.bloger.cz/IT-internet/Fuzzy-logika

[12] LEVINGS PP, BUNGERT J (March 2002). "The human beta-globin locus control region". Eur. J. Biochem. 269 (6): 1589–99. DOI:10.1046/j.1432-1327.2002.02797.x. PMID 11895428

[13] LOONEY, C, G., Interactive clustering and merging witha new fuzzy expected value, In: Pattern Recognition 35 (2002) 2413 – 2423, Computer Science Department 171, University of Nevada, Reno, NV 89557, USA 2001

[14] 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

[15] MOHYLOVÁ, J., KRAJČA V., Zpracování biologických signálů 1. vydání, Ediční středisko VŠB – TUO, 2006, 135s. ISBN 978-80-248-1491-9

[16] MURRAY, R., GRANNER, D., MAYES, P., RODWELL, V. Harperova biochemie. 4. české vyd. Jinočany: Nakladatelství H+H 2002. 53–64 s. ISBN 80-7319-013-3

[17] NASSER, Sara, Adrienne BRELAND, Frederick C. HARRIS a Monica NICOLESCU. A fuzzy classifier to taxonomically group DNA fragments within a metagenome. In: Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society, 2008.: NAFIPS 2008. New York, 2008, s. 1-6

[18] NAVARA, M., OLŠÁK, P., Základy fuzzy množin. 1. Vyd. Praha: Vydavatelství ČVUT, 2002. 136s. ISBN 80-01-02585-3

[19] NCBI. National center for biotechnology information. http://www.ncbi.nlm.nih.gov/, NIH, 2007

[20] Nehierarchické metody shlukování, [online], [cit. 2012-11-01]. Dostupné z WWW: http://is.muni.cz/th/172767/fi_b/5739129/web/web/nehiermet.html

[21] Obrázek dostupný z WWW: http://antranik.org/blood-components-hemoglobin-typerh-factor-agglutination/

[22] Obrázek dostupný z WWW: http://www.nature.com/nmeth/journal/v6/n4/images/nmeth.f.247-F3.jpg

[23] RACLAVSKÝ, Vladislav. Metody molekulární genetiky [online]. Ústav biologie Lékařské fakulty Univerzity Palackého v Olomouci, 2003, [cit. 2011-03-20]. Kapitola 8. Sekvenování DNA

[24] RATNASINGHAM, S. & HEBERT, P. D. N. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). Molecular Ecology Notes 7, 355-364. DOI: 10.1111/j.1471-8286.2006.01678.x

[25] SCHNEIDER TD (2002). "Consensus Sequence Zen". Appl Bioinformatics 1 (3): 111–119. PMC 1852464. PMID 15130839

[26] SCHNEIDER TD, STEPHENS RM (1990). "Sequence Logos: A New Way to Display Consensus Sequences". Nucleic Acids Res 18 (20): 6097–6100. DOI:10.1093/nar/18.20.6097. PMC 332411. PMID 2172928

[27] SOKAL R, MICHENER C (1958). "A statistical method for evaluating systematic relationships". University of Kansas Science Bulletin 38: 1409–1438

[28] TĚTHAL, J. Komparační analýza genomických dat pomocí grafické reprezentace. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2011, 54s. Vedoucí bakalářské práce Ing. Denisa Maděránková

[29] TOMANDL, J., TÁBORSKÁ, E. a kol. Biochemie I – Semináře. Brno: Masarykova univerzita, 2003. 12–16 s. ISBN 80-210-3056-9

[30] VUT, FEKT studijní materiály k předmětu Analýza biologických sekvencí (2011), garant předmětu: Provazník, I

[31] WEI YOU, KUN WANG, HUIXIAO LI, YANG JIA, XIAOQIN WU, YANING DU, Classification of DNA Sequences Basing on the Dinucleotide Compositions, Department of Mechanical and Electrical Engineering, North China Institute of Science and Technology 2009

[32] XU D, KELLER J M., POPESCU M, BONDUGULA R, Applications of Fuzzy Logic in Bioinformatics, University of Missouri-Columbia, USA, 2008 ISBN-10 1-84816-258-8

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

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

[35] ZADEH, L.A.: Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. IEEE Trans. Syst. Man. Cybern., 1, 1973, s. 28-44

[36] ZEMAN, D.: Aplikace procesu dolováni dat v biologii - genetice [online]. Brno: VUT, [2005]. [cit. 2012-11-01]. Dostupný z WWW: http://www.fit.vutbr.cz/study/courses/VPD/public/0405VPD-Zeman.pdf

[37] 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

[38] ZVÁROVÁ, Jana. Základy statistiky pro biomedicínské obory [online] 1. vydání. Praha: Karolinum, 1998. ISBN 80-7184-786-0. Dostupné také z: http://ucebnice.euromise.cz/index.php?conn=0&section=biostat1&node=5

[39] ŽVÁČKOVÁ, I, DNA Sekvenční motivy ovlivňující expresi transgenu, Brno: Masarykova univerzita, Přírodovědecká fakulta, 2012, 51 s

Seznam zkratek

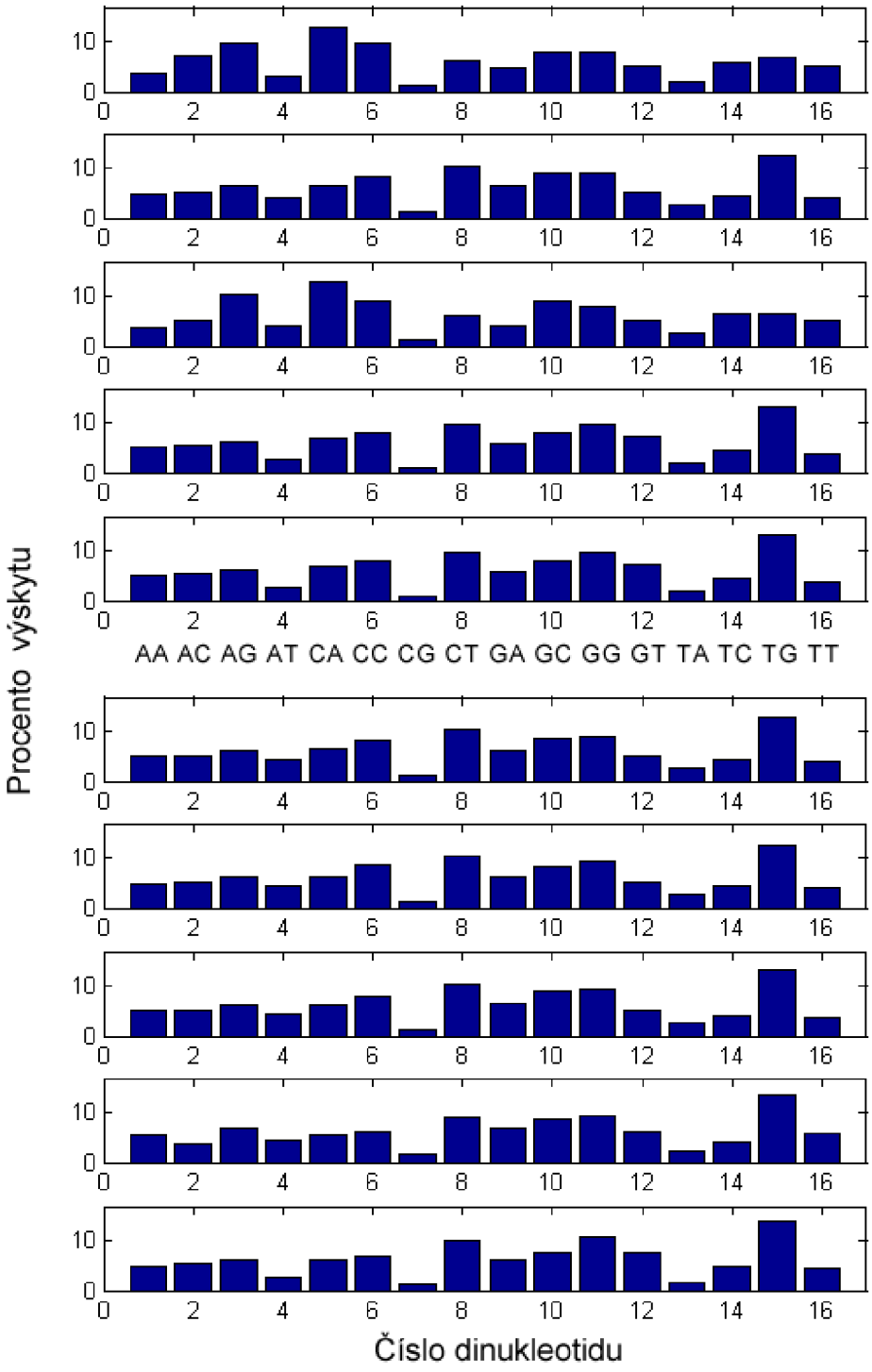
BOLD	The Barcode of Life Database
CO1	cytochrom oxydáza
DNA	deoxyribonukleová kyselina
EMBL	Evropská molekulárně biologická laboratoř
FMF	fuzzy funkce příslušnosti
FMF-JC	fuzzy funkce příslušnosti pro Jukes-Cantor
FMF-D	fuzzy funkce příslušnosti pro denzitu
JC	Jukes-Cantor

Seznam příloh

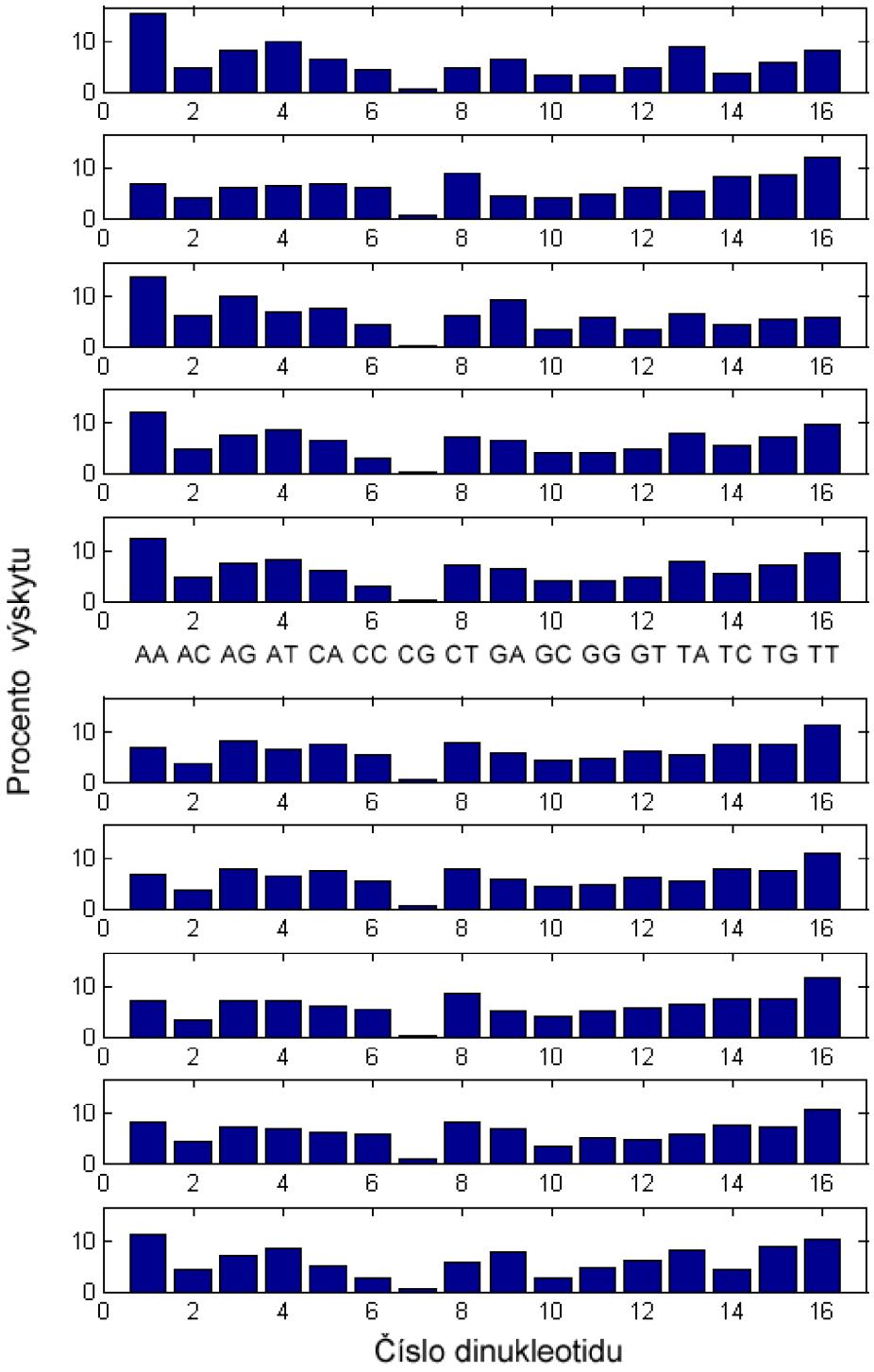
Seznam příloh na CD

Elektronická verze diplomové práce
Soubory analyzovaných sekvencí
Hlavní funkce a vnořené funkce

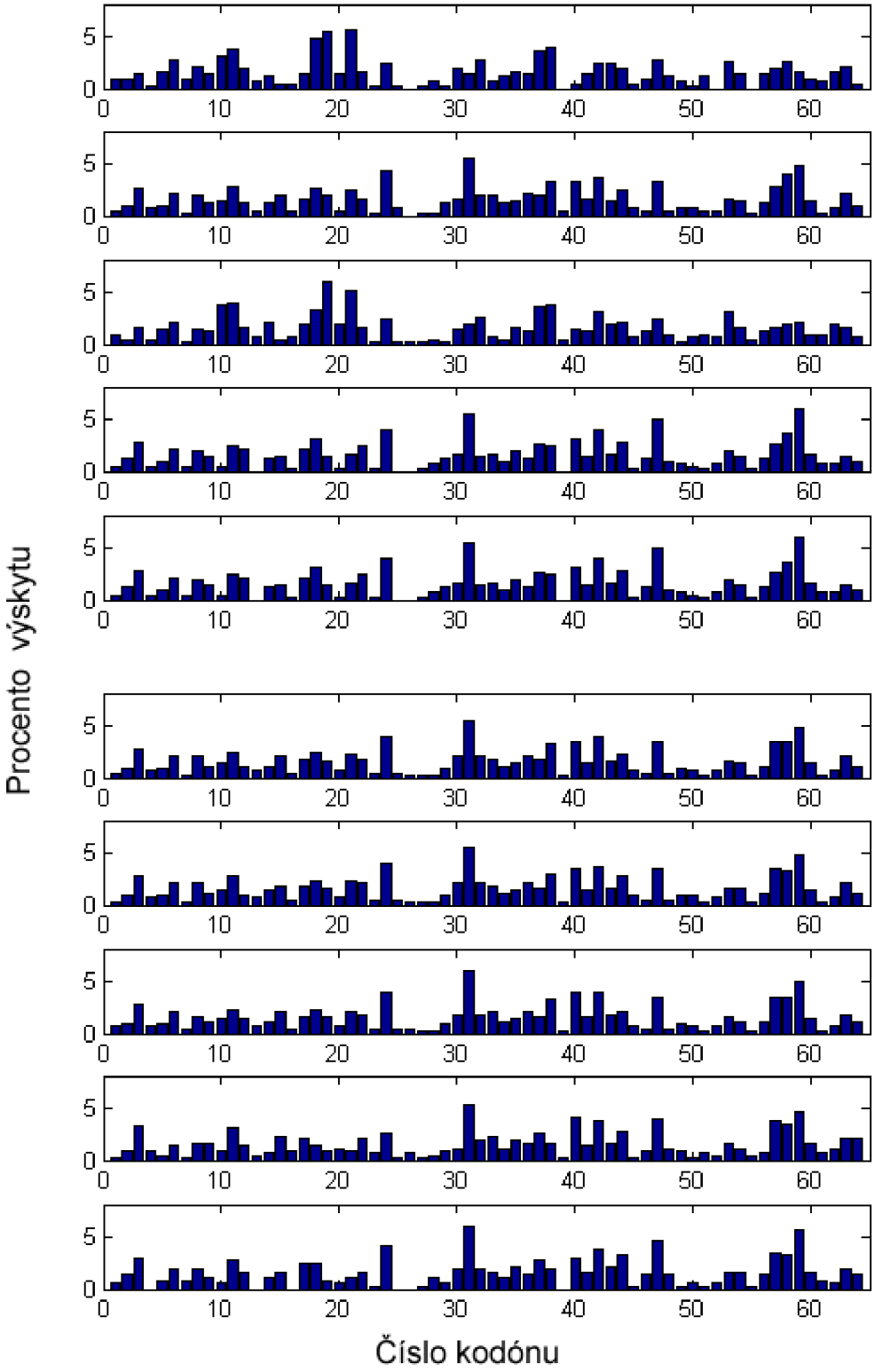
Příloha 1: Poměrný výskyt dinukleotidů v kódujících úsecích pro jednotlivé sekvence



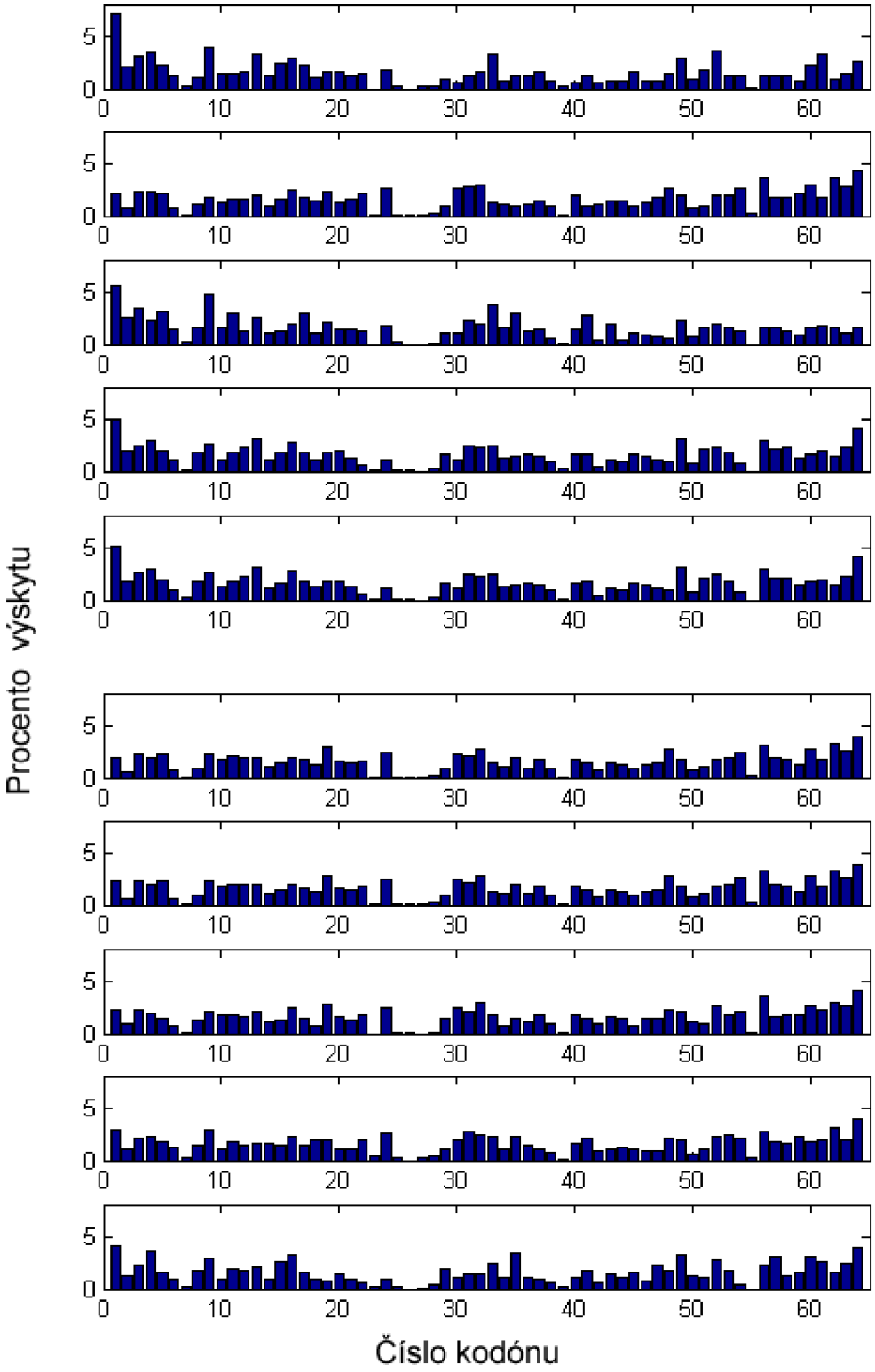
Příloha 2: Poměrný výskyt dinukleotidů v nekódujících úsecích pro jednotlivé sekvence



Příloha 3: Poměrný výskyt trinukleotidů v kódujících úsecích pro jednotlivé sekvence



Příloha 4: Poměrný výskyt trinukleotidů v nekódujících úsecích pro jednotlivé sekvence



Příloha 5: Hlavičky použitých sekvencí pro analýzu exonů a intronů

'gi|224589802:5246696-5248301 Homo sapiens chromosome 11, GRCh37.p10 Primary Assembly CDS(133..261,1112..1334,1465..1556)'
'gi|51763895:38287-54863 Mus musculus strain BALB/c chromosome 7 genomic contig, GRCm38.p1 alternate locus group BALB/c BALB/C_MMCHR7_CTG1 CDS(53..144,261..483,16321..16449) '
'gi|372099103:103812528-103813923 Mus musculus strain C57BL/6J chromosome 7, GRCm38.p1 C57BL/6J CDS(129..257,912..1134,1251..1342)'
'gi|71727226|gb|DQ126303.1| Homo sapiens isolate HbA-Ivc18 beta globin (HBB) gene, complete cds(3067..3158,3289..3511,4362..4490)'
'gi|71727222|gb|DQ126301.1| Homo sapiens isolate HbA-G37 beta globin (HBB) gene, complete cds(3067..3158,3289..3511,4362..4490)'
'gi|290756159|gb|GU057239.1| Mus musculus castaneus voucher IN47 beta-globin (Hbbt1) gene, Hbbt1-T1_A allele, complete cds(330..421,538..760,1413..1541)'
'gi|290756157|gb|GU057238.1| Mus musculus castaneus voucher IN45 beta-globin (Hbbt1) gene, Hbbt1-T1_B allele, complete cds(330..421,538..760,1415..1543)'
'gi|261873724|gb|GQ250395.1| Mus spretus beta-globin (Hbbt2) gene, Hbbt2-1 allele, complete cds(335..426,540..762,1417..1545)'
'gi|258513352:49022978-49024620 Bos taurus breed Hereford chromosome 15, Bos_taurus_UMD_3.1, whole genome shotgun sequence CDS (53..138,267..489,1390..1518)'
'gi|38226|emb|X02345.1| P.troglodytes beta-globin gene, exons 1-3 CDS(4189..4293,4412..4633,5484..>5532)'

Příloha 6: Rozdělení třídy Actinopterygii na jednotlivé řády

- Nadřád: chrupavčití (Chondrostei)
  - mnohoploutví (Polypteriformes)
  - jeseteři (Acipenseriformes)
- Nadřád: mnohokostnatí (Neopterygii)
  - kaprouni (Amiiformes)
  - kostlíni (Lepisosteiformes)
- Nadřád: kostnatí (Teleostei)
  - ostnojazyční (Osteoglossiformes)
  - tarponi (Elopiformes)
  - albulotvaří (Albuliformes)
  - holobřiší (Anguilliformes)
  - velkotlamky (Saccopharyngiformes)
  - bezostní (Clupeiformes)
  - maloústí (Gonorynchiformes)
  - máloostní (Cypriniformes)
  - trnobřiší (Characiformes)
  - sumci (Siluriformes)
  - nahohřbetí (Gymnotiformes)
  - štikotvární (Esociformes)
  - koruškotvární (Osmeriformes)
  - lososotvární (Salmoniformes)
  - velkoústí (Stomiiformes)
  - měkkorypí (Ateleopodiformes)
  - jinožábří (Aulopiformes)
  - hlubinovky (Myctophiformes)
  - leskyňovci (Lampridiformes)
  - vousatky (Polymixiiformes)
  - okounovci (Percopsiformes)
  - hrujovci (Ophidiiformes)
  - hrdloploutví (Gadiformes)
  - žabohlaví (Batrachoidiformes)
  - ďasové (Lophiiformes)
  - cípalové (Mugiliformes)
  - gavúni (Atheriniformes)
  - jehlotvární (Beloniformes)
  - halančíkovci (Cyprinodontiformes)
  - mořatky (Stephanoberyciformes)
  - pilonoši (Beryciformes)
  - pilobřiši (Zeiformes)
  - volnoostní (Gasterosteiformes)
  - hrdložábří (Synbranchiformes)
  - ropušnicotvární (Scorpaeniformes)
  - carouni (Gobiesociformes)
  - ostnoploutví (Perciformes)
  - platýsi (Pleuronectiformes)
  - čtverzubci (Tetraodontiformes)

Příloha 7: Funkce pro spočítání obsahu GC

```
function [rozlozeni,procento]=histCG(sekvence,delkaokna)

delka=length(sekvence); //spočítá délku sekvence

for i=1:(delka-delkaokna) //projíždí sekvenci pomocí plovoucího okna

    s=sekvence(i:i+delkaokna);

sumCG=sum(s=='C')+ sum(s=='G'); //spočítá výskyt C a G

procento(i)=(sumCG/delkaokna)*100;  //spočítá procento výskytu

end

rozlozeni=min(procento):0.5:max(procento); //vypočítá roztložení dat v
histogramu

end
```

Příloha 8: Funkce pro spočítání počtu dinukleotidů

```
function [dinukleotidy, pocetdinukleotidu, p_dinukleotidu]=pocetdinuk(seq)
acgt=('ACGT'); //zadá možnosti kódování nukleotidů
x=1;
p=1;
delka=length(seq); //spočítá délku sekvence
//vytvoření všech kombinací nukleotidů
for a=1:4

    nukleotid1=acgt(a); //vybírá postupně první ze 4 nukleotidů

    for b=1:4

        nukleotid2=acgt(b); //vybírá postupně druhý ze 4 nukleotidů

           celkove=[nukleotid1 nukleotid2]; //sestaví dinukleotid
             dinukleotidy(x,1:2)=celkove;
             pocetdinukleotidu(x,1)=length(strfind(seq,celkove)); //spočítá
četnost výskytu
             p_dinukleotidu(x,1)=pocetdinukleotidu(x,1)/delka*100; //spočítá
procento výskytu
             x=x+1;

    end
end

end
```

Příloha 9: Funkce pro spočítání počtu trinukleotidů

```
function [trinukleotidy, pocettrinuktrinukleotidu,
p_trinukleotidu]=pocettrinuk(seq)

acgt=('ACGT'); //zadá možnosti kódování nukleotidů
x=1;
p=1;
delka=length(seq); //spočítá délku sekvence
//vytvoření všech kombinací nukleotidů
for a=1:4

    kodon1=acgt(a); //vybírá postupně první ze 4 nukleotidů

    for b=1:4

        kodon2=acgt(b); //vybírá postupně druhý ze 4 nukleotidů
        for c=1:4

            kodon3=acgt(c);  //vybírá postupně třetí ze 4 nukleotidů
           celkove=[kodon1 kodon2 kodon3]; //sestaví trinukleotid

             trinukleotidy(x,1:3)=celkove;
             pocettrinuktrinukleotidu(x,1)=length(strfind(seq,celkove));
//spočítá četnost výskytu
             p_trinukleotidu(x,1)=pocettrinuktrinukleotidu(x,1)/delka*100;
//spočítá procento výskytu
             x=x+1;

        end
    end
end

end
```