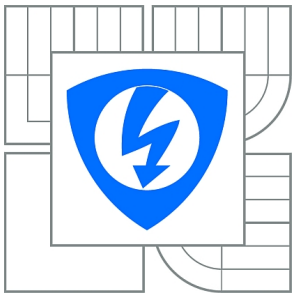
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

EVOLUČNÍ MODELY PRO VYHODNOCENÍ PŘÍBUZNOSTI ORGANISMŮ

EVOLUTIONARY MODELS FOR EVALUATION OF ORGANISMS RELATIONSHIP

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

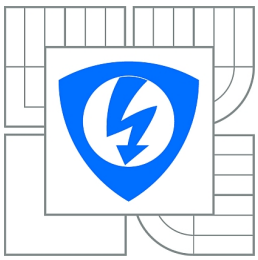
AUTOR PRÁCE — KATEŘINA GREGOROVÁ

AUTHOR

VEDOUCÍ PRÁCE — Ing. HELENA ŠKUTKOVÁ

SUPERVISOR

BRNO 2015



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Studentka: Kateřina Gregorová ***ID:*** 155573
Ročník: 3 ***Akademický rok:*** 2014/2015

NÁZEV TÉMATU:

Evoluční modely pro vyhodnocení příbuznosti organismů

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši metod pro vyhodnocení evoluční vzdálenosti DNA sekvencí a proteinových sekvencí v aminokyselinové a kodónové reprezentaci. 2) V programovém prostředí Matlab realizujte evoluční modely pro DNA sekvence. Výsledky pro více sekvencí prezentujte formou fylogenetického stromu. 3) Realizujte algoritmy pro stanovení evoluční vzdálenosti řetězců aminokyselin a kodónů. 4) Všechny realizované modely implementujte do společného programu s grafickým uživatelským rozhraním s možností grafického výstupu formou fylogenetického stromu. 5) Program otestujte na vhodně zvolených sekvencích z veřejných databází. Proveďte srovnání všech metod a výsledky diskutujte.

DOPORUČENÁ LITERATURA:

[1] HIGGS, PAUL G a TERESA K ATTWOOD. Bioinformatics and molecular evolution. Edtion ed.: Wiley-Blackwell, 2005. ISBN 978-1405106832.
[2] OTU, H. H. a K. SAYOOD. A new sequence distance measure for phylogenetic tree construction. Bioinformatics, Nov 2003, 19(16), 2122-2130.

Termín zadání: 9.2.2015 ***Termín odevzdání:*** 29.5.2015

Vedoucí práce: Ing. Helena Škutková
Konzultanti bakalářské práce:

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

UPOZORNĚNÍ:

ABSTRAKT

Práce je zaměřená na studium a popis evolučních modelů pro vyhodnocení evoluční vzdálenosti DNA sekvencí a proteinových sekvencí v aminokyselinové a kodónové reprezentaci. V rámci této práce byl vytvořen program, který vyhodnocuje genetickou vzdálenost sekvencí DNA a proteinů za použití některých evolučních modelů. Program vypočítá genetické vzdálenosti porovnávaných sekvencí a na jejich základě vykreslí fylogenetický strom. Takto lze relativně snadno a rychle vyhodnotit příbuznost organismů. Pro snadné ovládání je součástí vyhodnocovacího programu také grafické uživatelské rozhraní (GUI).

KLÍČOVÁ SLOVA

evoluce, mutace, evoluční model, substituce, substituční model, proteinová sekvence, rozdělení, metoda, fylogenetický strom

ABSTRACT

The work is focused on the study and description of evolutionary models for the evaluation of the evolutionary distances of DNA sequences and protein sequences in the amino acids and the codon representation. In the framework of this work was created a program that evaluates the genetic distance between DNA sequences and protein sequences for the use of some evolutionary models. The program calculates the genetic distance of the compared sequences and on the basis thereof renders the phylogenetic tree. This can be relatively easily and quickly evaluate the affinity of the organisms. For easy operation is part of the evaluation program also graphical user interface (GUI).

KEYWORDS

evolution, mutations, the evolutionary model, substitution, substitution model, protein sequences, the distribution, method, phylogenetic tree

GREGOROVÁ, K. *Evoluční modely pro vyhodnocení příbuznosti organismů.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2015. 41 s. Vedoucí bakalářské práce Ing. Helena Škutková.

PROHLÁŠENÍ

Prohlašuji, že svoji bakalářskou práci na téma Evoluční modely pro vyhodnocení příbuznosti organismů jsem vypracovala samostatně pod vedením vedoucí bakalářské 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 bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne 29. 5. 2015 .....

(podpis autora)

PODĚKOVÁNÍ

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

Dále děkuji své rodině, která mi při psaní této práce dopřála potřebný klid.

V Brně dne 29. 5. 2015 .....

(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

Úvod

Evoluce je dlouhodobý a samovolný proces, v jehož průběhu se rozvíjí a diverzifikuje pozemský život. Základními evoluční mechanismy jsou mutace, přirozený výběr a genetický drift. Všechny tyto procesy mají za následek diverzifikaci všech organismů. Bez těchto procesů by evoluce nebyla.

Věda, která se zabývá tímto evolučním vývojem, se nazývá fylogenetika. Dříve byla velmi subjektivní, avšak s objevením molekulární struktury DNA se stala vědou zcela objektivní, jelikož znaky na úrovni DNA jsou přesně definovatelné. Zároveň lze považovat genetické informace za číslicová data, která jsou vhodná k matematickému zpracování za využití výpočetní techniky.

K takovému matematickému zpracování slouží v dnešní době evoluční modely, které na základě určitých informací dokážou vypočítat genetickou vzdálenost organismů. Tyto modely se dělí do několika skupin a to na modely, které zpracovávají sekvence DNA v nukleotidové podobě, modely, které se zaměřují na výpočet genetické vzdálenosti z aminokyselinových sekvencí, nebo ze sekvencí kodónů.

Každá z těchto skupiny je odlišně výpočetně náročná. K těm nejjednodušším modelům patří právě ty, co zpracovávají jednotlivé nukleotidy, v podstatě porovnávají dvě sekvence a vyhledávají změny, které proběhly v jedné nebo druhé sekvenci. Úplně nejjednodušším modelem je JC69, který hodnotí veškeré změny. Komplikovanější je pak Kimurův dvouparametrický model, který odlišuje transice a transverze. S výpočetní složitostí však roste přesnost daného modelu.

Některé ze základních nukleotidových modelů se dají v pozměněné formě aplikovat také pro výpočet genetické vzdálenosti ze sekvence aminokyselin či kodónů.

1 MOLEKULÁRNÍ EVOLUCE A FYLOGENETIKA

V následujících kapitolách se budeme zabývat vyhodnocováním příbuznosti organismů na základě genetických dat. Pro pochopení je důležité vysvětlit základní pojmy a způsoby interpretace dosažených výsledků.

1.1 ZÁKLADNÍ EVOLUČNÍ MECHANISMY

Mezi základní evoluční mechanismy patří mutace, dále je to přirozený výběr a také genetický drift. Mutace jsou změny ve struktuře genetického materiálu respektující pravidla zápisu genetické informace. Jsou nezbytné pro biologickou evoluci. Podle fyzické povahy rozlišujeme bodové, řetězcové, chromozomové a genomové. Naším předmětem zájmu jsou především mutace bodové, na kterých evoluční modely závisí.

1.1.1 Bodové mutace

Bodové mutace spočívají nejčastěji ve změně jednoho nukleotidu za jiný, jedná se o tzv. záměnové mutace, substituce. Jestliže je nukleotid s určitým typem báze nahrazen nukleotidem s jiným typem báze, jedná se o transverzi, jestliže nukleotidem s bází stejného typu, jedná se o tranzici. Dále mezi bodové mutace patří delece a inzerce, při nichž se v určitém místě DNA mění počet nukleotidů, nejčastěji o jeden, poměrně časté jsou i dinukleotidové inzerce a delece.

Mutace můžeme rozdělit také podle jejich vlivu na strukturu proteinu. O synonymní mutaci hovoříme v případě, že se záměna nukleotidu nikam neprojevila na struktuře proteinu. To je možné proto, že stejná aminokyselina je kódována celou řadou různých proteinů. V opačném případě se jedná o mutace se změnou smyslu. Avšak pokud je aminokyselina nahrazena jinou, která má podobné fyzikálně-chemické vlastnosti, jedná se o záměnu konzervativní.

Dalším typem záměnových mutací jsou mutace nesmyslné. V těchto případech dochází k nahrazení kodónu některým z kodónů terminačních, tím pádem dojde při translaci k předčasnému ukončení syntézy proteinového řetězce. Tato změna e natolik drastická, že ve většině případů vzniká nefunkční protein. [1]

1.2 STRUKTURA A FUNKCE GENŮ

Jedná-li se o funkci, můžeme geny rozdělit do dvou skupin – protein-kódující geny a RNA-kódující geny. Protein-kódující jsou přepisovány do mRNA a následně přeloženy do sekvence aminokyselin proteinů. RNA-kódující geny jsou ty, které produkují tRNA, rRNA, snRNA a další.

Genetická informace nesená v nukleotidové sekvenci je nejprve přepsána do mRNA, která určuje pořadí aminokyselin v aminokyselinovém řetězci. Jednotlivé nukleotidy jsou

postupně čteny po trojicích – tzv. kodonech. Každý kodon je přeložen do určité aminokyseliny v polypeptidovém řetězci.

Genetický kód pro jaderné geny je s pár výjimkami univerzální pro prokaryota i eukaryota. Stejný genetický kód je používán i pro chloroplasty avšak mitochondriální geny používají mírně odlišný. Na tyto odlišnosti musíme dát při vyhodnocování příbuznosti pozor, jinak by došlo ke značnému zkreslení výstupních informací. Standartní genetický kód je zobrazen v tabulce (Tab. 1.1). [3]

Tab. 1.1: Standartní genetický kód

	T	**C**	**A**	**G**
T	TTT Phe (F)	TCT Ser (S)	TAT Tyr (Y)	TGT Cys (C)
	TTC	TCC	TAC	TGC
	TTA Leu (L)	TCA	**TAA Ter**	**TGA Ter**
	TTG	TCG	**TAG Ter**	TGG Trp (W)
C	CTT Leu (L)	CCT Pro (P)	CAT His (H)	CGT Arg (R)
	CTC	CCC	CAC	CGC
	CTA	CCA	CAA Gln (Q)	CGA
	CTG	CCG	CAG	CGG
A	ATT Ile (I)	ACT Thr (T)	AAT Asn (N)	AGT Ser (S)
	ATC	ACC	AAC	AGC
	ATA	ACA	AAA Lys (K)	AGA Arg (R)
	ATG Met (M)	ACG	AAG	AGG
G	GTT Val (V)	GCT Ala (A)	GAT Asp (D)	GGT Gly (G)
	GTC	GCC	GAC	GGC
	GTA	GCA	GAA Glu (E)	GGA
	GTG	GCG	GAG	GGG

2 MODELY EVOLUCE DNA

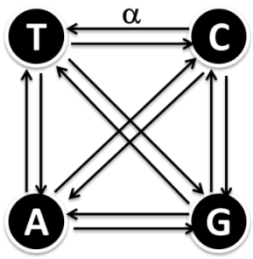
Evoluční modely nukleových kyselin a proteinů se používají ve fylogenetice jako základ definování evoluční vzdálenosti a k výpočtu pravděpodobností z množiny sekvencí. Konkrétní výpočet pravděpodobnosti je závislý na tom, co předpokládáme o charakteru jednotlivých potencionálních změn. Soubor těchto předpokladů tvoří tzv. substituční model.

2.1 PROPORCIONÁLNÍ VZDÁLENOST SEKVENCÍ

Nejjednodušším způsobem jak vyjádřit vztah mezi dvěma sekvencemi, je spočítat pozice ve kterých se liší, a tento počet vydělit celkovým počtem zkoumaných nukleotidů. Tato veličina se nazývá pozorovaná distance neboli p-distance. Tato metoda je velmi jednoduchá a intuitivní, ale problém nastává v případě, že s postupujícím časem se zvyšuje pravděpodobnost opakovaných substitucí. Například pokud dojde k substituci A →T a potom k substituci T→C, zachytíme pouze substituci A→C. V důsledku této chyby dochází k podhodnocení skutečné míry divergence mezi oběma sekvencemi. [2]

2.2 JUKESŮV – CANTORŮV MODEL (JC69)

Nejjednodušším substitučním modelem je jednoparametrický model JC69. Pravděpodobnost změny jednoho nukleotidu na jiný popisuje parametr α. Tento model předpokládá, že frekvence výskytu všech čtyř nukleotidů jsou totožné a pravděpodobnost změny kteréhokoliv nukleotidu v sekvenci na jiný je vždy stejná. Nukleotidové sekvence jsou tedy ekvivalentní.



Obr. 2.1: Model JC69

Ze stavového modelu (Obr. 2.1) lze vyvodit, že hodnota parametru α pro nukleotidové sekvence bude $\alpha = \frac{3}{4}$. Pro proteinové sekvence vycházející z 20 aminokyselin pak $\alpha = \frac{19}{20}$. Průměrný počet změn je dán součinem frekvence změn a času.

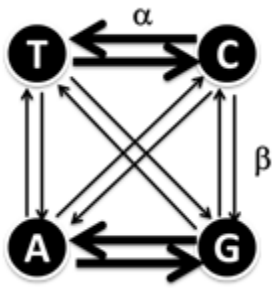
Pomocí vztahu 2.1 je možné vypočítat evoluční vzdálenost d bez znalosti času a frekvence výskytu změn. Hodnota p vyjadřuje relativní počet změn mezi sekvencemi, tedy hodnotu p-distance.

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \tag{2.1}$$

Jukes-Cantorův model dává přesné výsledky pouze tehdy, není-li množství změn mezi sekvencemi příliš velké. Pokud je hodnota p-distance vyšší než jedna je vhodné použít jeden ze složitějších modelů.[6]

2.3 KIMURŮV DVOUPARAMETROVÝ (K2P) A TŘÍPARAMETROVÝ (K3ST) MODEL

Model K2P je druhým nejjednodušším modelem. Oproti předchozímu modelu počítá s druhým parametrem β. Tyto parametry představují četnosti zvlášť pro transice a transverze. Běžně je v genetickém kódu poměr transicí a transverzí 1:2 což je znázorněno ve stavovém modelu na obrázku (Obr. 2.2). [5]



Obr. 2.2: Model K2P

U Kimurova modelu se zanedbává možný rozdílný počet pyrymidinů a purinů v sekvenci. Počet nukleotidů, které se mění v důsledku transice jsou označena P. Počet nukleotidů, které se mění v důsledku transverze jsou označena Q. Celkový počet míst, které se změní, je $p = P + Q$.

Evoluční vzdálenost vypočítáme pomocí následujícího vzorce:

$$d = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q). \tag{2.2}$$

Kimurův tříparametrový model se liší od předchozího tím, že rozlišuje celkem tři typy substitucí: transice, transverze A ↔ T, C ↔ G a transverze A ↔ C a T ↔ G. [5]

2.4 FALSENSTEINŮV (TAJIMŮV-NEIŮV) MODEL (F81)

Tento model již neuvažuje stejné frekvence pro všechny čtyři nukleotidy v sekvenci. Je to proto, že některé nukleotidy v sekvenci mohou být zastoupeny častěji než jiné a proto dochází u těchto nukleotidu k četnějším substitucím a naopak. Evoluční vzdálenost d se vypočítá následovně:

$$d = -\left(1 - (pA^2 + pC^2 + pG^2 + pT^2)\right) * log(1 - p\_dist/(1 - (pA\text{^}2 + pC\text{^}2 + pG\text{^}2 + pT\text{^}2))) \tag{2.3}$$

kde pA, pC, pG a pT jsou proporcionální četnosti jednotlivých nukleotidů a p je hodnota genetické vzdálenosti. [7]

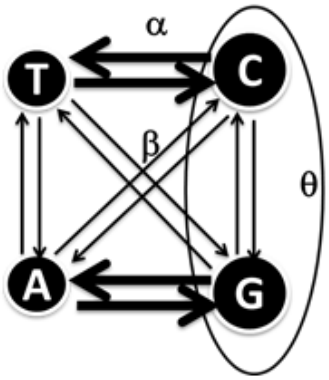
2.5 HASEGAWŮV – KISHINŮV – YANŮV MODEL (HKY85)

HKY85 model je obecnějším případem modelů K2P a F81, který pracuje kromě nestejného zastoupení jednotlivých bází v sekvencích také s rozdílnou rychlostí transicí a transverzí. Je to jeden z nejčastěji používaných substitučních modelů. [6] Vztah pro výpočet evoluční vzdálenosti je následující:

$$\begin{aligned} d = \; & -k1 * log(1 - pR/(2 * pA * pG) * P/2 - Q/(2 * pR)) \\ & -k2 * \log\left(1 - \frac{pY}{2 * pT * pC} * \frac{P}{2} - \frac{Q}{2 * pY}\right) - 2 * \\ & \left(pR * pY - pA * pG * \frac{pY}{pR} - pT * pC * \frac{pR}{pY}\right) * log(1 - Q/(2 * pR * pY)) \end{aligned} \tag{2.4}$$

2.6 TAMURŮV MODEL (T92)

Jedná se v podstatě o Kimurův dvouparametrický model rozšířený o třetí parametr θ_1. Tímto parametrem je proporcionální četnost G a C ve všech sekvencích. Tento parametr vychází z toho, že Guanin a Cytosin mezi sebou tvoří silnější vazbu (tři vodíkové můstky), kterou jde obtížněji narušit. Zvýšená proporcionální četnost G a C má tendenci tvořit stabilnější sekvenci. Pokud budeme uvažovat také proporcionální četnost A a T, můžeme tento model doplnit o čtvrtý parametr θ_2. Hodnota tohoto parametru potom bude $\theta_2 = 1 - \theta_1$. [9]



Obr. 2.3: Model T92

Evoluční vzdálenost se vypočítá následovně:

$$\begin{aligned} d = \; & -2 * th * (1 - th) * \log\left(1 - \frac{P}{2 * th * (1 - th)} - Q\right) \\ & -(1 - 2 * th * (1 - th)) * log(1 - 2 * Q)/2 \end{aligned} \tag{2.5}$$

2.7 TAMURŮV – NEIŮV MODEL (TN93)

Evoluční model, který rozšiřuje T92. Na rozdíl od HKY zohledňuje kromě transverzí také dva typy tranzicí. Transici mezi puriny A ↔ G a transici mezi pyrimidiny C ↔ T. Rovněž umožňuje použít různé četnosti jednotlivých nukleotidů. Jedná se o víceparametrický model (parametrem je α_1, α_2, β, π_A, π_C, π_G a π_T). [6],[8]

Evoluční vzdálenost je:

$$d = -\frac{2\pi_A\pi_G}{\pi_R}\ln\left(1-\frac{\pi_R}{2\pi_A\pi_G}P_1-\frac{1}{2\pi_R}Q\right)-\frac{2\pi_T\pi_C}{\pi_Y}\ln\left(1-\frac{\pi_Y}{2\pi_T\pi_C}P_2-\frac{1}{2\pi_Y}Q\right) - 2\left(\pi_R\pi_Y-\frac{\pi_A\pi_G\pi_Y}{\pi_R}-\frac{\pi_T\pi_C\pi_R}{\pi_Y}\right)\ln\left(1-\frac{1}{2\pi_R\pi_Y}Q\right) \quad (2.6)$$

Počet nukleotidů A a G, které se mění v důsledku transice je označen P_1. Počet nukleotidů C a T, které se mění v důsledku transice je označen P_2. Počet nukleotidů, které se mění v důsledku transverze je označen Q. [8]

2.8 ZOHLEDNĚNÍ IN/DEL MUTACÍ

Evoluční vzdálenost lze vypočítat také s ohledem na počet insertních či delečních mutací. Počet in/del mutací nezávisí na délce, tedy počtu skórovacích pomlček v mutaci.

$$d = \boldsymbol{ev.model} * \left(1-\frac{G}{G+L_S}\right)+\frac{G}{G+L_S}, \quad (2.7)$$

kde G = počet skórovacích pomlček v obou sekvencích, Ls = délka sekvencí bez pomlček.

Takto lze modifikovat každý vzorec pro výpočet evoluční vzdálenost ze substitucí aminokyselin i nukleotidů.[10] Modifikace Jukesova-Cantorova modelu pak bude následující:

$$d = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right) * \left(1-\frac{G}{G+L_S}\right)+\frac{G}{G+L_S} \quad (2.8)$$

3 MODELY EVOLUCE PROTEINOVÝCH SEKVENCÍ

Modely evoluce proteinových sekvencí patří mezi kvantitativní modely, které se zaměřují na výpočet evoluční vzdálenosti mezi aminokyselinami v sekvencích proteinů. K jejich vyhodnocování používáme dva odlišné přístupy a to výpočet substitucí aminokyselin v proteinové sekvenci a výpočet nukleotidové substituce v kodonech.

3.1 AMINOKYSELINOVÉ SUBSTITUČNÍ MODELY

Základním prvkem těchto modelů je výpočet proporcionální vzdálenosti neboli *p*-distance, která se vypočítá jako podíl počtu změněných pozic *n* a délky sekvence bez mezer *L*. Budeme-li uvažovat následující zarovnané sekvence výsledná hodnota *p*-distance je rovna $0,\overline{4}$. [10]

```
K V L R D N I - G
K - J R D N R Q G
  *   *     *   *
```

Obr. 3.1: Aminokyselinové sekvence

$$p = \frac{n}{L} = \frac{4}{9} = 0,\overline{4} \tag{3.1}$$

Stejně jako při výpočtu genetické vzdálenosti pro sekvence nukleotidu, se jedná pouze o výpočet konečných stavů, proto musíme pro výpočet evoluční vzdálenosti aplikovat pravděpodobnostní modely pro popis řídkých jevů. Těmito modely rozumíme binomické, Poissonovo či Gama rozdělení. [10]

3.1.1 Poissonovo rozdělení

Poissonovo rozdělení bývá označováno jako rozdělení řídkých jevů, neboť se podle něj řídí počet výskytů události (jevů), které mají velmi malou pravděpodobnost výskytu. Pokud je hodnota *p*-distance menší než 0,3, rozložení nemá smysl a výsledky jsou totožné. Evoluční vzdálenost se vypočítá následovně: [20]

$$d = -\ln(1-p), \tag{3.2}$$

kde p udává hodnotu p-distance.

3.1.2 Gama rozdělení

U gama rozdělení počítáme ještě s dalším parametrem γ, který řídí evoluční rychlost. Obor hodnot tohoto parametru je $\langle 0{,}65; 3\rangle$.

$$d = \gamma\left[(1-p)^{-\frac{1}{\gamma}} - 1\right] \tag{3.3}$$

Pokud je hodnota *p*-distance menší než 0,2 toto rozložení opět nedává smysl a výsledek je roven přímo *p*-distanci. [13]

3.1.3 Jukesův-Cantorův model pro aminokyseliny

Tento model byl původně odvozen pro bílkoviny. Vychází ze stejného základu jako JC model pro nukleotidy avšak hodnoty 4 a 3 označující počet různých nukleotidů a počet typů změn na jiný nukleotid, byly nahrazeny čísly 20 a 19, které odpovídají situaci u aminokyselin. Stejně jako Jukes-Cantorův model pro nukleotidy tento model nepředpokládá různou substituční rychlost různých aminokyselinových substitucí ani vliv frekvence aminokyselin v proteinu, tedy ochotu používat určitou aminokyselinu. Evoluční vzdálenost se pak vypočítá pomocí následujícího vzorce: [4]

$$d = -\frac{19}{20}\ln\left(1-\frac{20}{19}p\right) \tag{3.4}$$

Obecněji lze vzorec napsat pomocí parametru B, který udává pravděpodobnost změny na jiný znak.

$$d = -B\ln\left(1-\frac{p}{B}\right) \tag{3.5}$$

Jukesův-Cantorův model lze upravovat pomocí již zmíněných korekcí.

3.1.4 Jukesův-Cantorův model s Gama korekcí

JC model pro aminokyseliny je již v základní podobě používán s Poissonovou korekcí. Gama korekce se používá pro delší evoluční dobu a větší rychlost změn u zkoumaných sekvencí. Gama korekce je vhodná jak pro aminokyseliny, tak i pro základní nukleotidový model. [10]

Vzorec pro výpočet evoluční vzdálenosti má následující podobu:

$$d = B.\gamma\left[\left(1-\frac{p}{B}\right)^{-\frac{1}{\gamma}}-1\right], \tag{3.6}$$

kde B = 3/4 pro nukleotidy a 19/20 pro aminokyseliny.

3.2 MODELY KODÓNOVÉ SUBSTITUCE

Modely, které jsou zde doposud prezentovány, předpokládají, že jednotlivé nukleotidy (aminokyseliny) jsou vzájemně nezávislé, to znamená, že neberou v úvahu strukturu kodonu. Jelikož je každá aminokyselina kódována tzv. nukleotidovým tripletem (kodonem) a každý triplet může být složen ze čtyř nukleotidových bází (A, C, G, T), existuje 4 x 4 x 4 = 64 možných kodonů, z nichž 61 kóduje konkrétní aminokyselinu a 3 jsou tzv. terminační (stop) kodony. Jelikož je genetický kód degenerovaný, to znamená, že většina z 20 aminokyselin může být kódována více než jedním kodonem, některé nukleotidové substituce nepovedou k záměně výsledné aminokyseliny. Substituce mající tuto vlastnost se nazývají synonymní, v opačném případě se jedná o nesynonymní substituce. Pokud při změně jednoho nukleotidu vznikne terminační kodon, dochází k substitucím neplatným.[2], [19]

CTT (Leu) → CTC (Leu) → TTG (Leu) – synonymní mutace

AGC (Ser) → AGG (Arg) → TGG (TRP) – nesynonymní mutace

TAT (Tyr) → TAA (ter.) – neplatná substituce

3.2.1 Metoda Nei-Gojobori (1986)

Z tabulky nukleotidových tripletů (Tab. 1.1) vidíme, že všechny substituce nukleotidu na druhé pozici v kodonu jsou nesynonymní, tedy v jejich důsledku dochází k nahrazení celé aminokyseliny. Avšak část substitucí, ke kterým dochází na první a třetí pozici v kodonu jsou synonymní. Za předpokladu stejné frekvence všech nukleotidů a jejich náhodného střídaní je tato část 5 % na první pozici a 72 % na pozici třetí. [11]

Počet synonymních změn s a počet nesynonymních změn n umíme vypočítat zvlášť pro každý kodon. Označíme-li počet synonymních substitucí na ité pozici daného kodonu f_i (i = 1, 2, 3), potom s a n lze vypočítat následovně:

$$s = \sum_{i=1}^{3} f_i \tag{3.7}$$

$$n = 3 - s \tag{3.8}$$

Například, v případě kodonu TTA (Leu), $f_1 = 1/3$ (T → C), $f_2 = 0$ a $f_3 = 1/3$ (A → G), potom s = 2/3 a n = 7/3.

Pro sekvence DNA složené z r kodonů je celkový počet synonymních a nesynonymních změny vyjádřen jako:

$$S = \sum_{j=1}^{r} S_i \tag{3.9}$$

$$n = 3r - S \tag{3.10}$$

V porovnání dvou sekvencí se vždy užívají průměrné hodnoty S a N.

Pro výpočet počtu synonymních a nesynonymních nukleotidových mutací u páru homologních sekvencí, porovnáváme obě sekvence po jednotlivých kodonech a počítáme počet nesynonymních a synonymních mutací u každého kodonového páru zvlášť. V případě, že došlo k substituci pouze u jednoho kodonu, můžeme hned rozhodnout, zda došlo k synonymní či nesynonymní mutaci.

Například, v případě porovnání kodonového páru GTT (Val) a GTA (Val) víme, že se jedná o mutaci synonymní. Označíme-li počet synonymních mutací s_d a počet nesynonymních mutací n_d, potom v tomto příkladu bude $s_d = 1$ a $n_d = 0$.

Jestliže se v porovnávání kodonového páru vyskytují dvě substituce, například při srovnání TTT a GTA, existují dvě různá řešení. První řešení je následující: TTT (Phe) ↔ GTT (Val) ↔ GTA (Val), zde došlo k jedné nesynonymní a jedné synonymní substituci. Ve druhém řešení TTT (Phe) ↔ TTA (Leu) ↔ GTA (Val), tady došlo ke dvěma nesynonymním mutacím. Jestliže budeme předpokládat, že oběma možnostem dochází se stejnou pravděpodobností, potom $s_d = 0{,}5$ a $n_d = 1{,}5$. [11]

Pokud dojde ke třem substitucím v porovnání dvou kodonů, existuje šest různých řešení a v každém řešení jsou tři mutační kroky. Pokud bychom porovnaly tato řešení s řešením v případě dvou substitucí, dostaneme stejnou výslednou hodnotu s_d a n_d. Evoluční cestu mezi synonymními a nesynonymními distancemi stanovíme jako permutaci všech změněných nukleotidů bez opakování: [11]

$$P\,(k) = k!, \tag{3.11}$$

k - počet změn, ke kterým došlo v jediném kodonu.

Nyní je jasné, že celkový počet synonymních a nesynonymních substitucí může být získán sečtením těchto hodnot u všech kodonů:

$$S_d = \sum_{j=1}^{r} s_{dj} \tag{3.12}$$

$$N_d = \sum_{j=1}^{r} n_{dj}, \tag{3.13}$$

kde s_{dj} a n_{dj} jsou hodnoty s_d a n_d jtého kodonu a r je počet porovnávaných kodónových párů.

Pomocí následujících rovnic můžeme tedy odhadnout podíl synonymních a nesynonymních mutací:

$$p_S = \frac{S_d}{S}, \tag{3.14}$$

$$p_N = \frac{N_d}{N}, \tag{3.15}$$

kde S a N je průměrný počet synonymních a nesynonymních mutací v porovnání dvou sekvencí.

Pro odhad počtu synonymních substitucí d(S) a nesynonymních substitucí d(N) na daném místě můžeme použít Jukesův-Cantorův vzorec pro výpočet evoluční vzdálenosti:

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right), \tag{3.16}$$

kde p je rovno p_S, nebo p_N. [11]

Př. 1: Uvažujme dva kodony TTT a GTA. U těchto kodonů došlo ke dvěma změnám, proto platí, že počet možných cest je roven:

$$P\,(2) = 2! = 2 \tag{3.17}$$

1. TTT (Phe) → GTT (Val) → GTA (Val)
2. TTT (Phe) → TTA (Leu) → GTA (Val)

1. $s_{d(1)} = 1;\ n_{d(1)} = 1$
2. $s_{d(2)} = 0;\ n_{d(2)} = 2$

Potom je celková distance rovna:

$$s_d = \frac{s_{d(1)} + s_{d(2)}}{2} = 0{,}5 \tag{3.18}$$

$$n_d = \frac{n_{d(1)} + n_{d(2)}}{2} = 1{,}5 \tag{3.19}$$

Př. 2: Uvažujeme-li dva kodony, které se změnily na třech pozicích, pak počet cest je roven:

$$P\,(3) = 3! = 6 \tag{3.20}$$

TTG (Leu) → AGA (Arg)

TTG (Leu) → ATG (Met) → AGG (Arg) → AGA (Arg)

TTG (Leu) → ATG (Met) → ATA (Ile) → AGA (Arg)

TTG (Leu) → TGG (Trp) → AGG (Arg) → AGA (Arg)

TTG (Leu) → TGG (Trp) → TGA (ter.) → AGA (Arg)

TTG (Leu) → TTA (Leu) → ATA (Ile) → AGA (Arg)

TTG (Leu) → TTA (Leu) → TGA (ter.) → AGA (Arg)

Jelikož se cesty s terminačním kodonem nepočítají, je výsledná hodnota $s_d = 3/4$ a $n_d = 9/4$. [10]

3.2.2 Modifikovaná metoda Nei-Gojobori

Jelikož dosavadní postup počítá hodnotu všech možných synonymních substitucí S s rovnocennou pravděpodobnostní transitních i transverzních změn, je vhodné metodu Nei-Gojobori modifikovat. Metodu modifikujeme tak, že zavedeme parametr R, který je roven poměru počtu transicí a transverzí:

$$\frac{\alpha}{\alpha + 2\beta} = \frac{R}{1 + R}, \tag{3.21}$$

kde α a β jsou relativní četnosti transicí a transverzí.

Jestliže R = 0,5, potom bude tato metoda shodná s metodou Nei-Gojobori. Jestliže R > 0,5, potom je počet synonymních míst menší, než udávala původní metoda a naopak počet nesynonymních míst bude větší. [3]

Celkový počet substitucí v sekvencích s parametrem R se vypočítá následujícím způsobem:

$$S_R = \sum_{j=1}^{r} s_{Rj} \tag{3.22}$$

$$N_R = \sum_{j=1}^{r} n_{Rj}, \tag{3.23}$$

kde s_{Rj} a n_{Rj} jsou hodnoty s_R a n_R jtého kodonu a r je počet porovnávaných kodónových párů.

Pro změnu kodonu TTT → TTC, kde je jedna synonymní substituce na 3. pozici, bude hodnota S_R rovna:

$$S_R = 0 + 0 + \frac{R}{1 + R} \tag{3.24}$$

Obdobně pro nesynonymní substituce TTT → CTT, kde došlo k jedné nesynonymní substituci na 1. pozici:

$$N_R = \frac{R}{1 + R} + 0 + 0 \tag{3.25}$$

Proporcionální distance se vypočítají následovně:

$$p_S = \frac{S_d}{S_R}, \tag{3.26}$$

$$p_N = \frac{N_d}{N_R}, \tag{3.27}$$

3.2.3 Li-Wu-Luo metoda

U této metody rozeznáváme tři třídy i = {0, 2, 4}, které popisují substituci kodónu. Jestliže každá substituce určité pozice je nesynonymní, nebo nesmyslná, potom tyto kodony řadíme do

třídy L_0. Do této skupiny patří všechny substituce, které proběhnou na druhé pozici u všech kodonů, většina změn na pozici první a také změny třetí pozice ATG (Met) a TGG (Trp).

Pokud alespoň jedna substituce dané pozice patří mezi synonymní, potom tyto kodony řadíme do třídy L_2. Dohromady do této třídy patří 24 kodonů, kde dochází k substitucím na třetí pozici, např. CAY (His), kde Y = T nebo C, tři leucinové kodony, kde dochází k alespoň jedné synonymní substituci na pozici první, např. YAR, kde R = A nebo G a čtyři argininové kodony CGR a AGR. Do této skupiny zahrnujeme také substituce, které probíhají na třetích pozicích isoleucinových kodonů.

Jestliže jsou všechny možné substituce daného místa synonymní, potom tyto kodony řadíme do třídy L_4. Celkem do této třídy řadíme 32 kodonů, kde dochází k substituci na třetí pozici, např. GTN (Val), kde N je jakákoliv báze. Počet míst v těchto třídách spočítáme v každé ze dvou porovnávaných sekvencí a výsledné průměrné číslo je pak výsledná hodnota L_0, L_2, L_4.

Účelem těchto klasifikací, je odhadnout míru synonymních a nesynonymních substitucí odděleně. Ve třídě L_0 dochází pouze k nesynonymním a ve třídě L_4 pouze k synonymním substitucím. U třídy L_2 vedou všechny transverze k nesynonymním a transice k synonymním změnám.

Pro výpočet stanovíme ve všech třídách proporcionální četnost P_i (transice) a Q_i (transverze) tak, že porovnáváme kodónové sekvence kodon po kodonu a aproximujeme podle Kimurova modelu [12]:

$$A_i = -\frac{1}{2} ln(1 - 2P_i - Q_i) + \frac{1}{4} ln(1 - 2Q_i) \tag{3.28}$$

$$B_i = -\frac{1}{2} ln(1 - 2Q_i) \tag{3.29}$$

Podle tohoto modelu pak stanovíme evoluční distance:

$$d_S = \frac{3[L_2 A_2 + L_4(A_4 + B_4)]}{L_2 + 3L_4} \tag{3.30}$$

$$d_N = \frac{3[L_0(A_4 + B_4) + L_2 B_2]}{3L_0 + 2L_2} \tag{3.31}$$

4 POUŽITÁ DATA

Základem výpočtu evoluční vzdálenosti organismů je odhalení homologních částí porovnávaných sekvencí. Homologní části, jsou takové, které vznikly odvozením od společného předka. Potřebujeme najít takové ortologní geny, abychom mohli změnu znaku za jiný považovat za substituci a záměnu znaku za mezeru za inserci, případně deleci. S tímto předpokladem pracují všechny evoluční modely.

Data používaná v této bakalářské práci jsou získaná z veřejné databáze GenBank [14] dne 22. 5. 2015. Testování programu jsem prováděla na souborech sekvencí vybraného proteinu. Primárně jsem zvolila HBB (hemoglobin beta) gen, který obsahuje pokyny pro výrobu bílkoviny zvané beta-globin. Mezi další vybrané geny patří ALB (albumin), CRH (Corticotropin-releasing hormone), STAR (Steroidogenic acute regulatory protein) a NEFL (Neurofilament light polypeptide). Důležitým kritériem pro výběr sekvencí byla jejich délka, byla jsem limitována výpočetní náročností programu.

Pro analýzu za pomoci nukleotidových modelů jsem vybrala 16S rRNA mitochondriální gen, který byl již v minulosti považován za velmi vhodný gen k fylogenetickým analýzám a to proto, že na tomto genu dochází k pomalému evolučnímu vývoji. Získaná data, však neobsahují translační přepis, proto nejsou vhodná pro výpočet distancí za použití aminokyselinových a kodónových modelů.

Soubory sekvencí obsahují vždy data od několika organismů tak, abychom byli schopni interpretovat získané výsledky a srovnat jednotlivé modely.

5 REALIZACE ALGORITMŮ EVOLUČNÍCH MODELŮ

Dříve než budeme analyzovat výstupy jednotlivých modelů, je vhodné popsat funkce navržených algoritmů a zároveň ověřit jejich správnost. Všechny tyto algoritmy tvoří komplexní program pro výpočet evoluční vzdálenosti organismů. Pro usnadnění a zpřehlednění je vytvořeno grafické uživatelské rozhraní, které bude popsáno v následujících kapitolách.

Správnost modelů ověříme na uměle vytvořených sekvencích sek1 a sek2.

```
sek1 = ACCTCAGCAACC

sek2 = ACGAGATTTACC
```

5.1 ALGORITMUS PRO VÝPOČET EVOLUČNÍ VZDÁLENOSTI DLE NUKLEOTIDŮ

Primárním parametrem všech modelů je výpočet p-distance. Hodnota p-distance u uvedených sekvencí p = 0,5. Vypočítá se jako součet všech rozdílných pozic dělený délkou sekvencí.

Dalšími parametry pro výpočet patří proporcionální četnost transicí P a transverzí Q, četnosti jednotlivých nukleotidů P_i (i = A, C ,T ,G) četnosti purinů P_R a pyrimidinů P_Y a theta, což je proporcionální četnost C a G. Výpočetní vzorce a výsledné hodnoty pro modelové sekvence jsou shrnuty v následující tabulce (Tab. 5.1).

Tab. 5.1: Evoluční modely pro výpočet evoluční vzdálenosti dle sekvencí nukleotidů

Model	**Vzorec**	**Výsledná hodnota**
JC69	$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$	0,8240
K2P	$d = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q)$	0,8857
F81	$d = -\left(1 - (pA^2 + pC^2 + pG^2 + pT^2)\right) * log(1 - p\_dist/(1 - (pA\text{^}2 + pC\text{^}2 + pG\text{^}2 + pT\text{^}2)))$	0,8710
T92	$d = -2 * th * (1 - th) * \log\left(1 - \frac{P}{2*th*(1-th)} - Q\right) - (1 - 2 * th * (1 - th)) * log(1 - 2 * Q)/2$	0,8857
TN93	$d = -\frac{2\pi_A\pi_G}{\pi_R}\ln\left(1 - \frac{\pi_R}{2\pi_A\pi_G}P_1 - \frac{1}{2\pi_R}Q\right) - \frac{2\pi_T\pi_C}{\pi_Y}\ln\left(1 - \frac{\pi_Y}{2\pi_T\pi_C}P_2 - \frac{1}{2\pi_Y}Q\right) - 2\left(\pi_R\pi_Y - \frac{\pi_A\pi_G\pi_Y}{\pi_R} - \frac{\pi_T\pi_C\pi_R}{\pi_Y}\right)\ln\left(1 - \frac{1}{2\pi_R\pi_Y}Q\right)$	0.9603
HKY	$d = -k1 * log(1 - pR/(2 * pA * pG) * P/2 - Q/(2 * pR)) - k2 * \log\left(1 - \frac{pY}{2 * pT * pC} * \frac{P}{2} - \frac{Q}{2 * pY}\right) - 2 * \left(pR * pY - pA * pG * \frac{pY}{pR} - pT * pC * \frac{pR}{pY}\right) * log(1 - Q/* pR * pY))$	0.9356

Tento algoritmus je v programovém prostředí realizován funkcí evolDNA jejímiž vstupy jsou porovnávané sekvence a výstupy hodnota evoluční vzdálenosti pro jednotlivé modely.

5.2 ALGORITMUS PRO VÝPOČET EVOLUČNÍ VZDÁLENOSTI DLE AMINOKYSELIN

Primárním výpočetním parametrem je opět hodnota p-distance. Tyto evoluční modely jsou v principu postaveny na modelu JC69. Zde došlo pouze ke změně hodnoty frekvence záměn z 3/4, jak je tomu u nukleotidových modelů na 19/20. Tento model má pak podobu:

$$d = -\frac{19}{20}\ln\left(1 - \frac{20}{19}p\right) \tag{5.1}$$

Aby nedocházelo ke zkreslení výsledů, je na model aplikována Poissonova či Gamma korekce. Programová realizace je popsána funkcí evolAK.

U uvedených algoritmů je také možno pomocí funkce indel zohlednit výskyt insercí či delecí v porovnávaných sekvencích. Výstupem této funkce je parametr Rg. Model JC69 se zohledněním in/del mutací vypadá následovně:

$$d = -B \ln\left(1 - \frac{p}{B}\right) * (1 - Rg) + Rg \qquad (5.2)$$

5.3 ALGORITMUS PRO VÝPOČET EVOLUČNÍ VZDÁLENOSTI DLE KODONŮ

Tento algoritmus je ze zde uvedených nejsložitější, jelikož v něm figuruje několik jednotlivých funkcí. Komplexní funkce nese název KODONY, jejímiž vstupy jsou opět porovnávané sekvence, ale také navíc proměnná gen_kod, která určuje, s jakým typem genu budeme pracovat. Ovládání tohoto nastavení je v GUI.

Výpočet evoluční vzdálenosti dle kodonů je realizován metodou Nei-Gojobori a je jediným algoritmem, který ve výstupu odděluje synonymní a nesynonymní mutace, tedy takové mutace kdy dochází nebo nedochází ke změně smyslu aminokyseliny, která je daným kodonem kódována. Samotný výpočet probíhá v několika fázích.

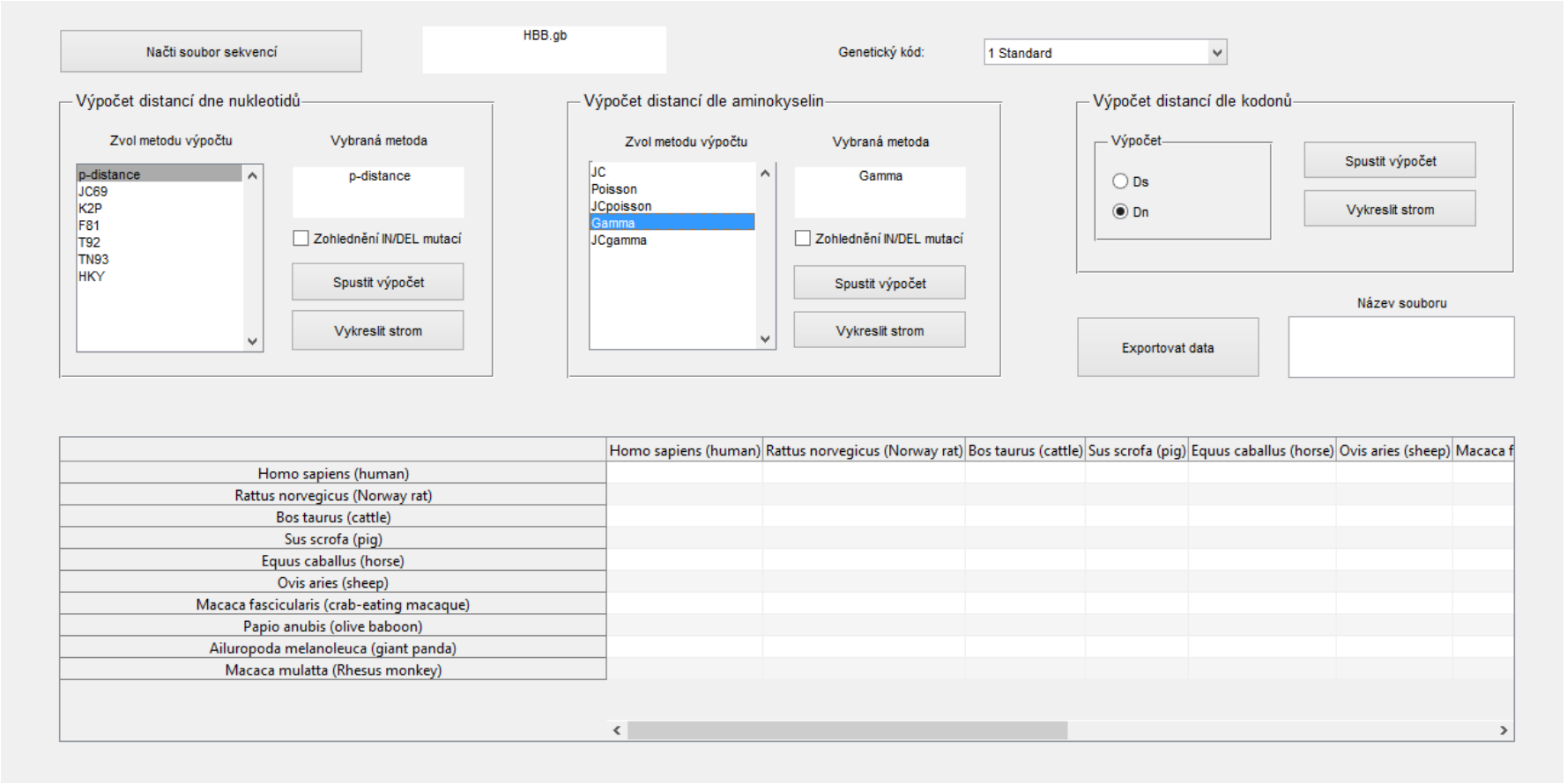
První fází je výpočet synonymních a nesynonymních substitucí, ke kterým došlo v jednotlivých sekvencích. Funkce S1 a S2 určují počet synonymních změn v první a druhé sekvenci a podobně je vypočítán i počet změn nesynonymních (N1 a N2). Výsledná hodnota S a N je dána průměrem hodnot S1,S2 a N1,N2.

Vstupy funkce SdNd jsou již jednotlivě separované kodony z porovnávaných sekvencí. Tato funkce nejen že hodnotí ke kolika synonymním a nesynonymním změnám došlo, ale určuje také na které pozici. Výstupem jsou pak proměnné Sd a Nd.

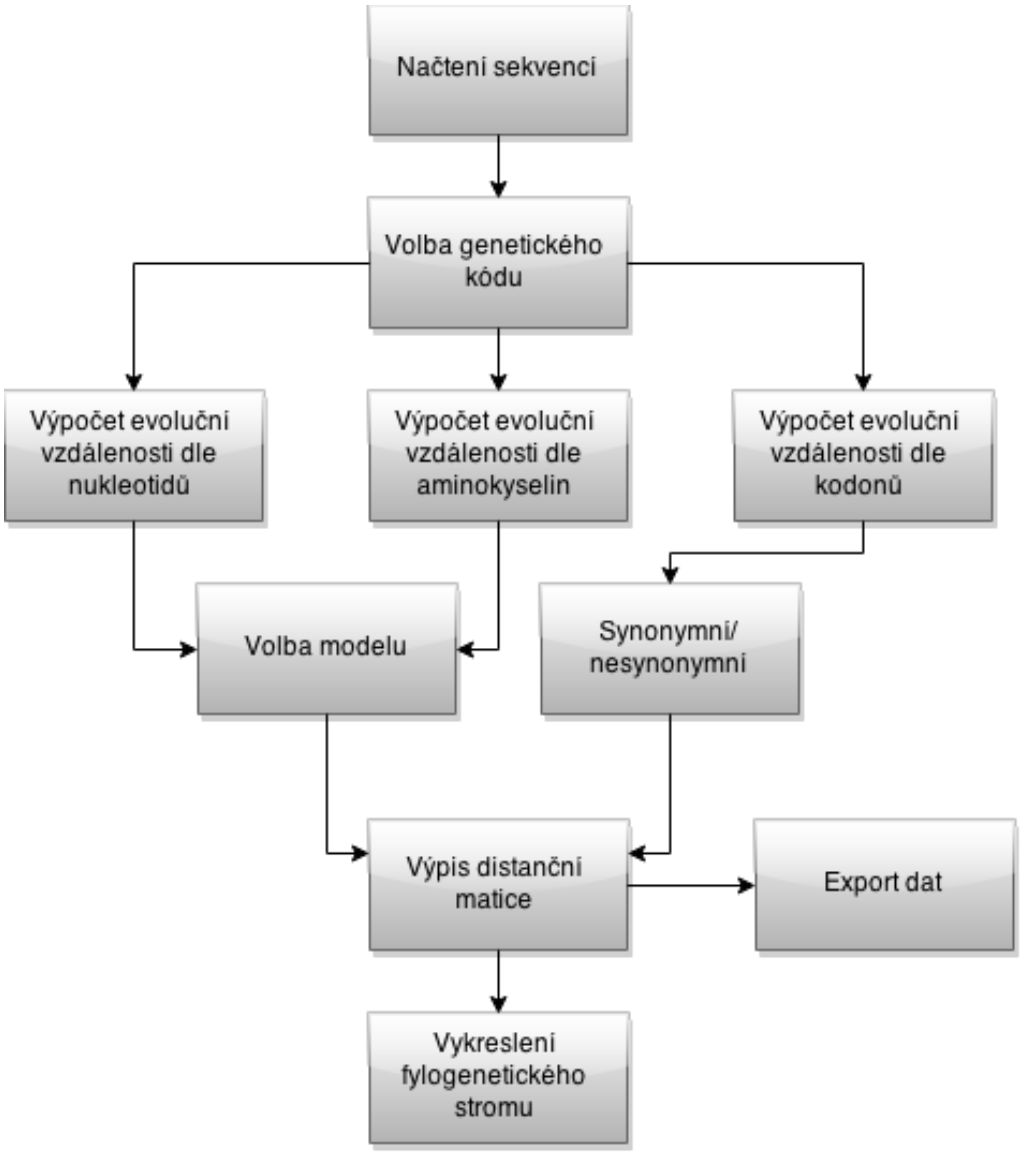
Všechny tyto subfunkce jsou volány v hlavní funkci KODONY. Ke konečnému výsledku se dostaneme tak, že použijeme opět Jukesův-Kantorův model a místo proporcionální vzdálenosti p počítáme s hodnotou ps nebo pn. (Viz vztah 3.26 a 3.27)

5.4 GRAFICKÉ UŽIVATELSKÉ ROZHRANÍ

Do grafického rozhraní jsou implementovány všechny výše popsané algoritmy tak, aby byly uživatelsky přístupné, přehledné a snadno ovladatelné. Interface programu je znázorněn na obrázku 5.1 (Obr. 5.1). Schéma propojení jednotlivých ovládacích prvků je znázorněno blokovým schématem na obrázku 5.2 (Obr. 5.2).



Obr. 5.1: Interface programu



Obr. 5.2: Blokové schéma programu

Načtení sekvencí

Po spuštění programu musíme nejprve načíst soubor sekvencí v GenBank formátu. Po stisknutí tlačítka „Načti soubor sekvencí“ se otevře nové okno, které umožnuje vybrat kterýkoliv soubor daného formátu uložený v počítači. Po okamžitém načtení sekvencí se v dolní části okna zobrazí tabulka se jmény organismů, jejichž geny následně porovnáváme.

Volba genetického kódu

Uživatel má možnost výběru genetického kódu. Je zde určitý předpoklad znalosti typu porovnávaných sekvencí.

Panely pro výpočet distancí

Program obsahuje tři nezávislé panely pro výpočet evoluční vzdálenosti. V každém panelu je možnost volby modelu, tlačítko pro spuštění výpočtu a tlačítko pro vykreslení stromu. Panely pro nukleotidy a aminokyseliny navíc obsahují možnost zohlednění in/del mutací. Vždy ale musí být zachována posloupnost výpočet a až poté vykreslení stromu.

Export dat

Tlačítko export dat umožňuje uložit hodnoty s distanční matice ve formátu.xls (Microsoft Excel). Do editovacího pole uživatel napíše, pod jakým názvem chce hodnoty uložit. Tato funkce je přínosem zejména tehdy, když potřebujeme s daty dále pracovat.

6 ANALÝZA EVOLUČNÍCH MODELŮ

Evoluční modely představené a realizované v této práci jsou základními hodnotícími prvky fylogenetické analýzy. Jejich vývoj začal již v roce 1965, kdy pánové Zuckerkandl E. a Pauling L. navrhli teorii molekulárních hodin, která říká, že rychlost molekulární evoluce je v průběhu času přibližně konstantní pro všechny proteiny ve všech liniích.

V teoretickém popisu evolučních modelů však není uvedeno na jaký typ dat je vhodné jednotlivé modely aplikovat, jaké jsou rozdíly ve výstupech a proč k těmto odlišnostem dochází. V této kapitole jsem proto využila realizované programové rozhraní a další metody pro vyhodnocování fylogenetické analýzy k řešení těchto problémů.

6.1 FYLOGENETICKÉ STROMY

Fylogenetické stromy představují grafické znázornění příbuzenských vztahů mezí různými taxonomickými jednotkami, o nichž lze předpokládat, že mají společného předka. Příbuzenské vztahy se posuzují podle morfologické či genetické podobnosti. Místo taxonomických jednotek, lze hodnotit také příbuznost jednotlivých organismů jako je to v této práci.

Z fylogenetických stromů lze určit evoluční vzdálenost pomocí délky větví, to ale závisí na volbě konstrukční metody stromu. Některé metody konstrukce fylogenetických stromů tuto vzdálenost zanedbávají. Tyto metody nejsou pro hodnocení výsledků v této práci ideální. Proto jsem pro vykreslení fylogenetických stromů zvolila metodu spojování sousedů (neighbor-joining).

6.1.1 Neighbor-joining

Tato metoda, pracuje na iterativním postupu, při němž v každém kroku spojuje dva sousedy v jeden uzel. Provádí shlukování výběrem sousedních sekvencí, jejichž součet délek větví vůči ostatním sekvencím je nejmenší. Tato metoda začíná u hvězdicovitého stromu, kde jsou spojeny všechny OTU (Operační taxonomické jednotky) v jeden centrální uzel. Poté vybere dvojici

OTU s nejmenší distancí a vypočítá sumu délek větví od jejich společného uzlu ke zbylému hvězdicovému uspořádání. Rekurzivně se tak tvoří v každém kroku nový strom pro menší množinu objektů. [15], [17]

6.2 HODNOCENÍ PODOBNOSTI STROMŮ

K hodnocení vzájemné podobnosti fylogenetických stromů se běžně využívá několik metod. V této práci jsem použila dvě z nich a to výpočet Pearsonova korelačního koeficientu a výpočet Robinson-Fouldovy vzdálenosti.

6.2.1 Pearsonův korelační koeficient

Korelační koeficient slouží ke stanovení závislosti mezi dvěma veličinami – zde pro stanovení závislosti distancí vypočítaných pomocí různých modelů. Nejjednodušším vztahem dvou metrických proměnných je vztah lineární. Tento koeficient nabývá hodnot v intervalu (-1; 1), kde hranice -1 označuje nepřímou lineární závislost a hranice 1 přímou lineární závislost. [16]

6.2.2 Robinson- Fouldova vzdálenost

Základem je obsahové porovnání shluků dvou stromů. Postupně takto můžeme vyhodnotit nejpřesnější strom vůči zvolenému referenčnímu. Robinsonovu -Fouldovu vzdálenost (dále jen R-F vzdálenost) získáme ze vzorce

$$d_{r,f} = \frac{(n_{c1-c2} + n_{c2-c1})}{2 * n}, \tag{6.1}$$

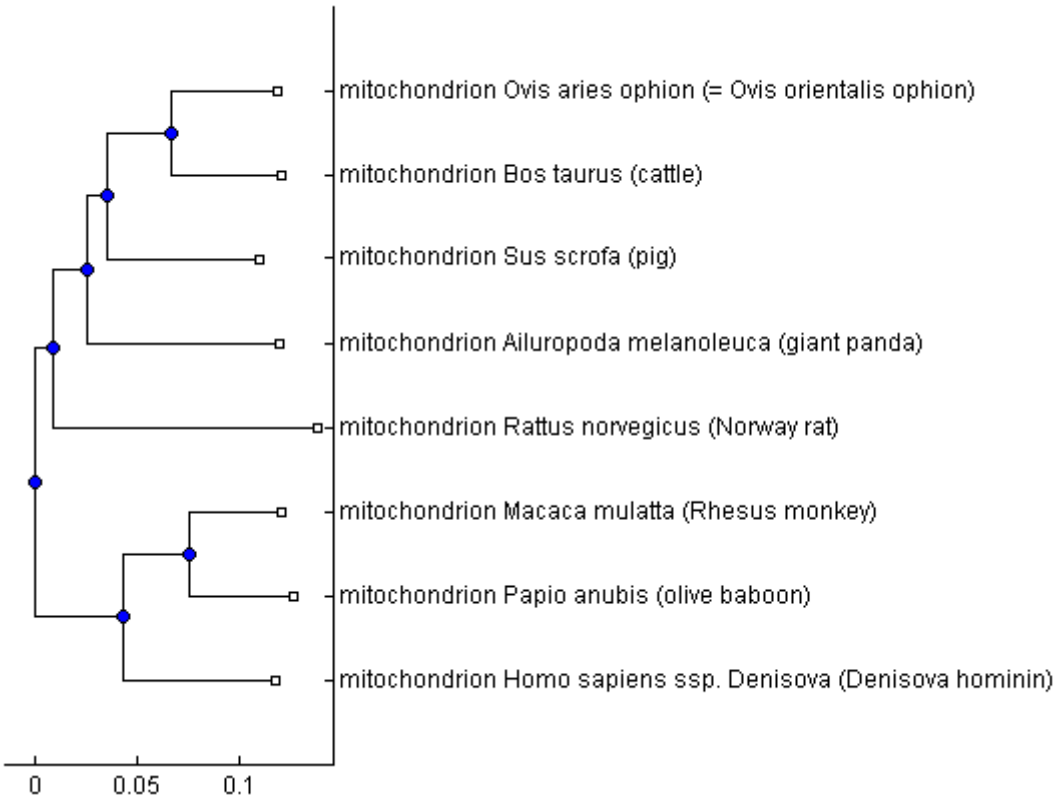
kde n_{c1-c2} vyjadřuje počet shluků vyskytujících se pouze v prvním stromu, n_{c2-c1} počet shluků vyskytujících se pouze v druhém stromu a n je počet shluků u obou sekvencí. [17]

7 ANALÝZA NUKLEOTIDOVÝCH SUBSTITUČNÍCH MODELŮ

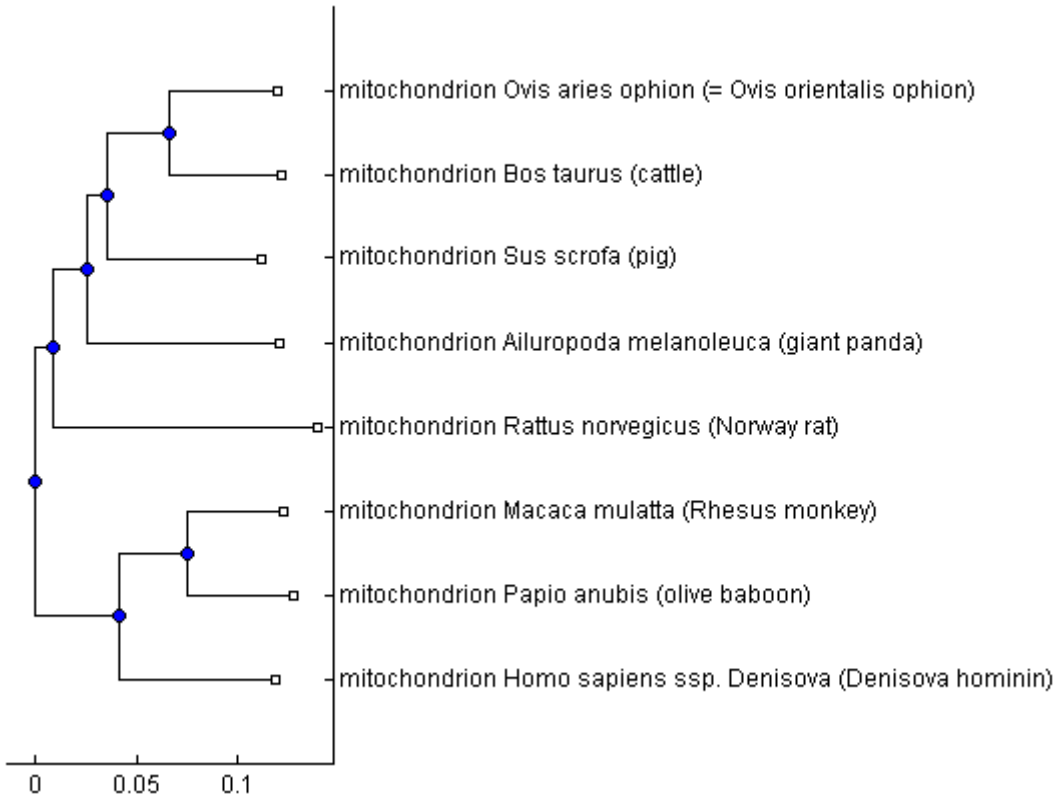
Pro zhodnocení dosažených výsledků u nukleotidových modelů jsem zvolila dva konkrétní modely a to nejjednodušší JC69 a model, který počítá s nejvíce parametry TN93. Vstupními daty je soubor sekvencí genu 16s rRNA.

Z topologie stromů je patrné, že výběr modelu nemá na strukturu stromu žádný vliv. U těchto sekvencí dochází k minimální diverzifikaci, proto je hodnota dosaženého Pearsonova koeficientu R = 0,998. V případě Robinson-Fouldovy vzdálenosti je $d_{r,f} = 0$.

Stejný pokus jsem provedla i s HBB proteinem a výsledky byly obdobné R = 0,999 a $d_{r,f} = 0$.



Obr. 7.1: Fylogenetický strom modelu JC69 pro gen 16s rRNA



Obr. 7.2: Fylogenetický strom modelu TN93 pro gen 16s rRNA

8 ANALÝZA AMINOKYSELINOVÝCH SUBSTITUČNÍCH MODELŮ

U aminokyselinových substitučních modelů jsem primárně porovnávala vliv Poissonovy a gamma korekce u Jukes-Cantorova modelu. K porovnávání jsem opět použila HBB protein.

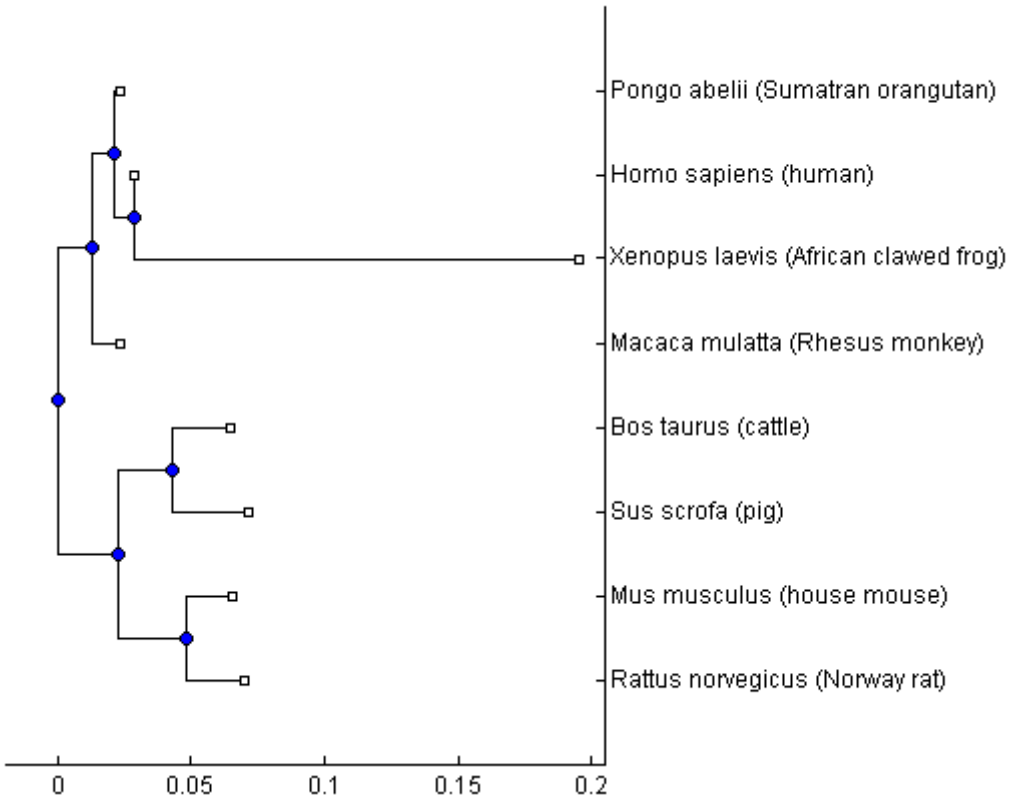
Výsledné hodnoty korelačního koeficientu pro porovnání samotného modelu a modelu s Poissonovou korekcí R = 1, model s korelační koeficient modelu s gamma korekcí R = 0,999. Z těchto informací vyplývá, že se jedná o velmi podobné sekvence, u kterých korekce nemají velký vliv. Stejně tak topologie fylogenetických stromů se nezměnila ani u jednoho modelu.

K zásadnějším změnám topologie stromu však došlo při použití proteinu NEFL a porovnání Jukes-Cantorova modelu a Jukes-Cantorova modelu s gamma korekcí. Zde však došlo s největší pravděpodobností k výskytu artefaktů.

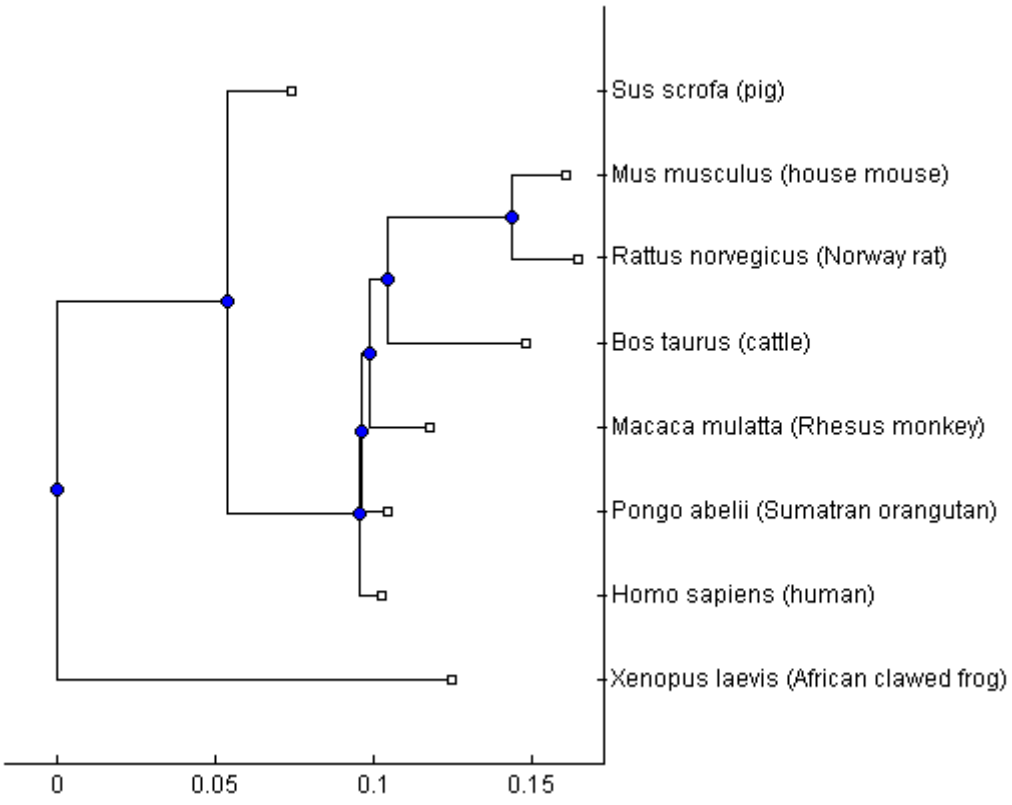
8.1.1 Artefakty v topologii stromu

Jedná-li se o velmi vzdálené sekvence, mohou se vytvářet tzv. artefakty dlouhých větví. Tyto artefakty bývají způsobeny výrazně odlišnou substituční rychlostí [10]

- Přitahování dlouhých větví: (LBA – long branch attraction) dvě velmi rozdílné sekvence (mezi sebou i vůči ostatním) jsou přitahovány k sobě a směrem ke kořeni
- Odpuzování dlouhých větví: (LBR – long branch repulsion) dva příbuzné geny se např. vlivem selekčních tlaků výrazně pozměnily a jsou odpuzovány od sebe
- Vyrušování dlouhých větví: (LBD – long branch distract) jedna dlouhá větev ovlivňuje topologii celého stromu



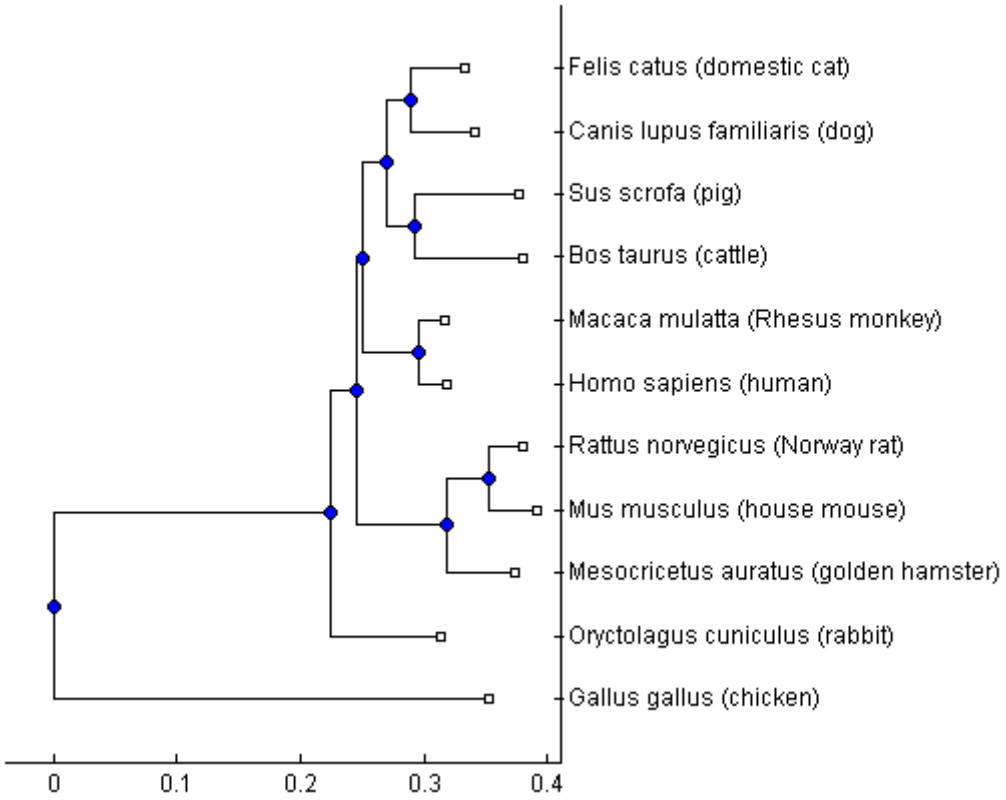
Obr. 8.1: Fylogenetický strom modelu JC pro protein NEFL



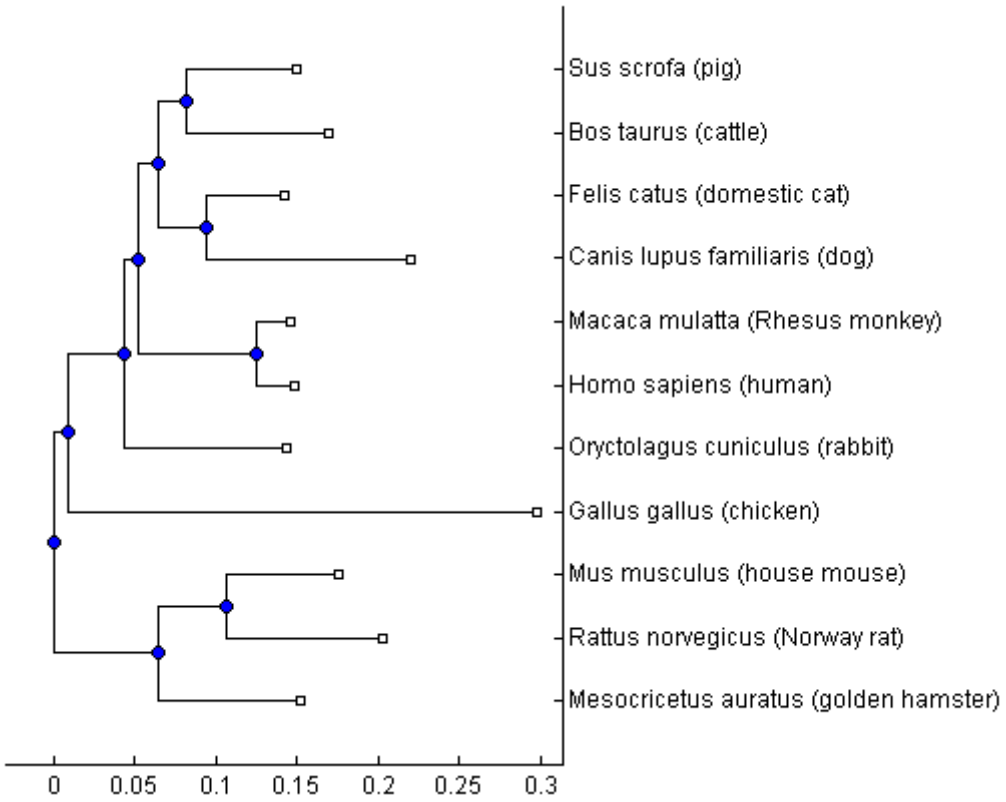
Obr. 8.2: Fylogenetický strom modelu JC s gamma korekcí pro protein NEFL

9 ANALÝZA KODÓNOVÝCH SUBSTITUČNÍCH MODELŮ

K analýze kodónových substitučních modelů jsem využila protein albumin. Kodónové substituční modely počítají zvlášť vzdálenost na základě synonymních a nesynonymních mutací, jak bylo uvedeno v předchozích kapitolách. V porovnání fylogenetických stromů modelu je tato odlišnost značně patrná. Hodnota Pearsonova koeficientu je pro tyto dvě metody R = 0,8101. A hodnota Robinson-Faudovy vzdálenosti $d_{r,f}$ = 0,2222.



Obr. 9.1: Fylogenetický strom hodnot Dn pro ALB protein



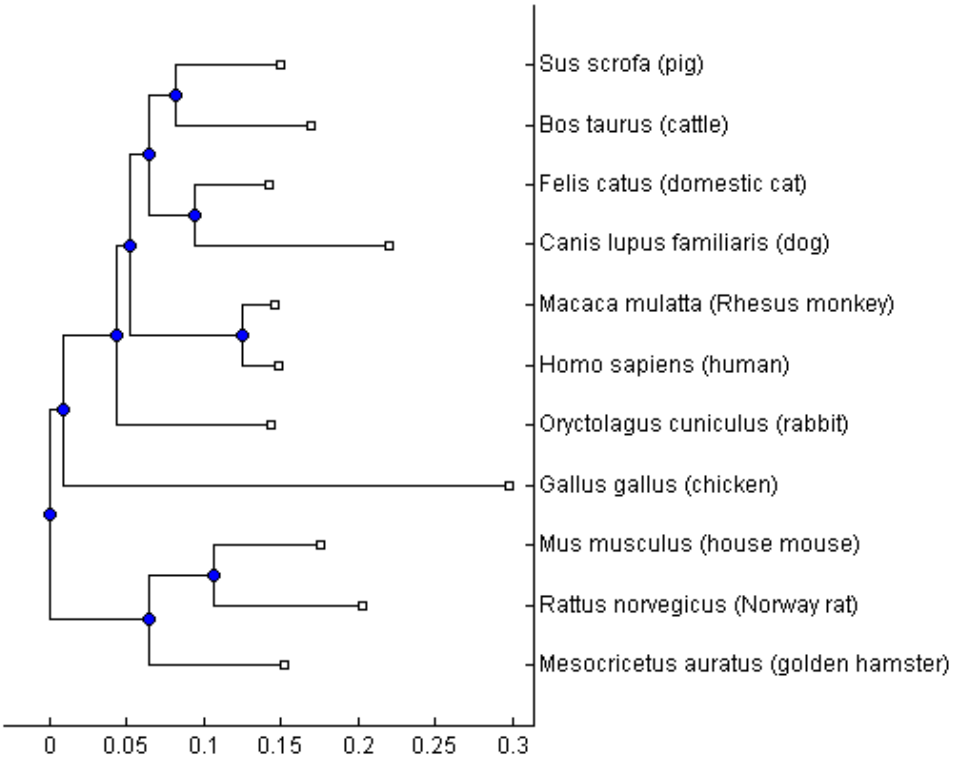
Obr. 9.2: Fyloenetický strom hodnot Ds pro ALB protein

10 KOMPLEXNÍ ZHODNOCENÍ POUŽITÝCH MODELŮ

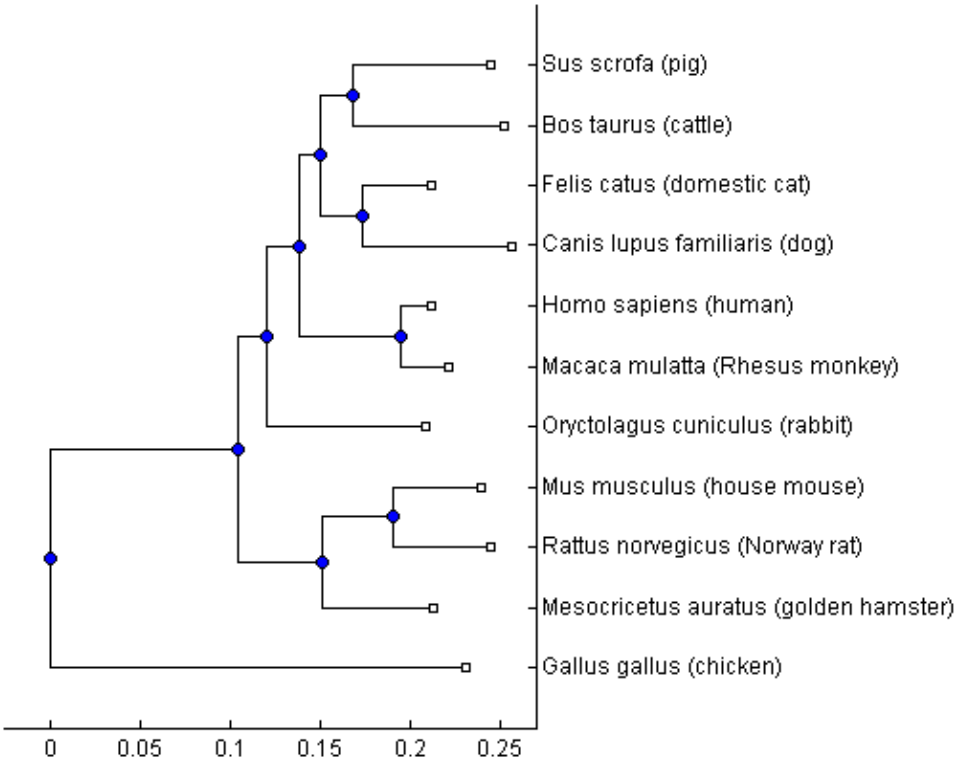
Zde se nabízí otázka, zda je lepší analyzovat sekvence podle proteinových či nukleotidových sekvencí. Optimální volba závisí na úrovni evolučního vztahy vyšetřovaných sekvencí. Pokud spolu sekvence úzce souvisí, pak bude DNA analýza pravděpodobně úspěšnější, protože umožňuje detekci synonymních změn. Pokud jsou studovány hlubší evoluční vztahy, pak je naopak výhodnější analýza proteinových sekvencí, protože ke změnám v proteinových sekvencích dochází pomaleji. To souvisí i s výstupními hodnotami dn a ds u výpočtu evoluční vzdálenosti na základě kodónů. Hodnota dn, tedy hodnota vypočítaná na základě nesynonymních mutací bude bližší výsledné hodnotě evoluční vzdálenosti u aminokyselinových modelů. Naopak hodnota ds bude více podobná evoluční vzdálenosti nukleotidovým modelům.

Pro praktické zhodnocení všech popsaných modelů jsem využila opět ALB protein kódující gen. Jelikož se neprokázal zásadní vliv výběru jednotlivých modelů v rámci hodnocení dle nukleotidů, aminokyselin a kodonů, je vhodné porovnat vlivy výběru modelu také mezi těmito skupinami. Pro tento účel jsem zvolila modely Tamura-Nei, Jukes-Cantor s gamma korekcí a obě varianty pro hodnocení dle nukleotidů.

Modely jsem porovnávala podle předpokládaných vlastností a to Tamura-Nei spolu s Ds a Jukes-Cantor s gamma korekcí s Dn.



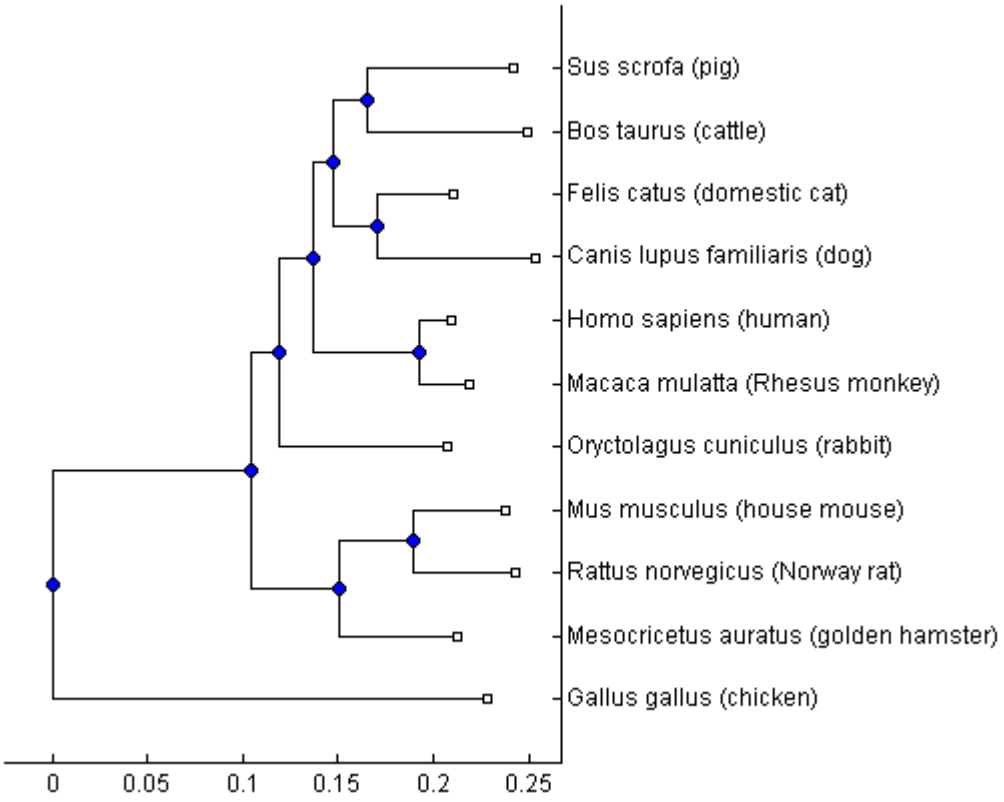
Obr. 10.1: Fylogenetický strom modelu Ds pro ALB protein



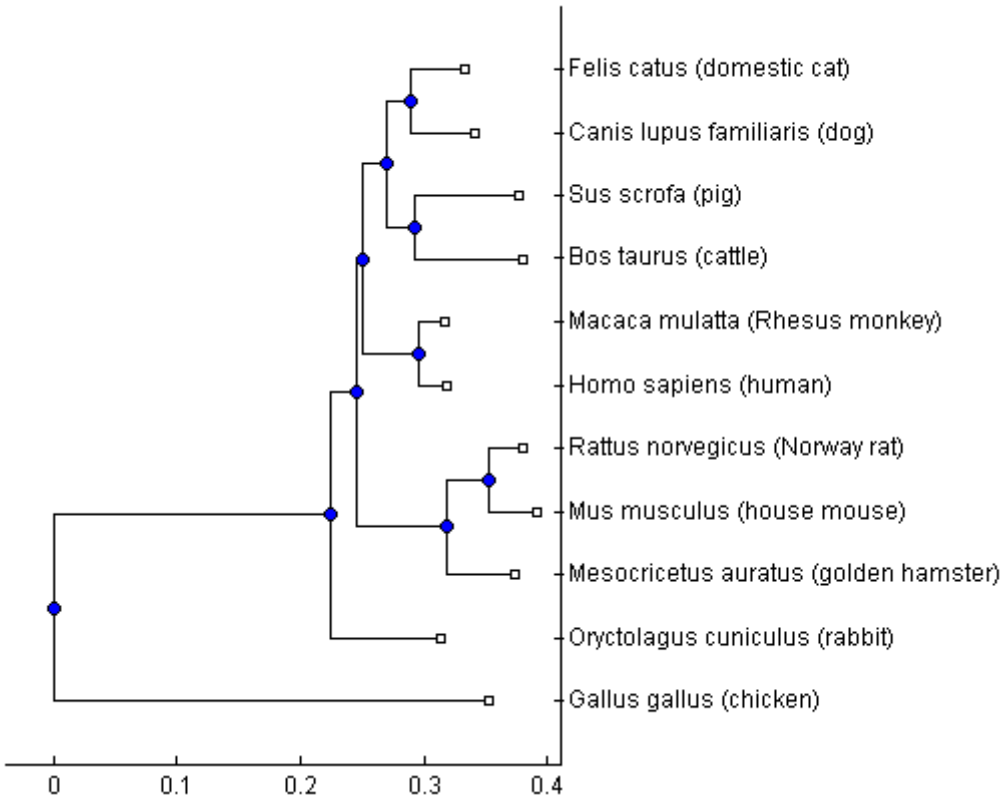
Obr. 10.2: Fylogenetický strom modelu TN93 pro ALB protein

Hodnota korelačního koeficientu pro tyto dva modely R = 0,9203 a Robinsonova-Faundova vzdálenost $d_{r,f}$ = 0.1111.

V dalším porovnání jsou modely Jukes-Cantor s gamma korekcí a Dn. Zde R = 0,9650 a $d_{r,f}$ = 0.



Obr. 10.3: Fylogenetický strom modelu JC s gamma korekcí pro ALB protein



Obr. 10.4: Fylogenetický strom modelu Dn pro ALB protein

Všechny předpoklady byly potvrzeny. Na obrázcích je vidět topologie porovnávaných fylogenetických stromů a je patrné, že se liší jak v délce větví tak v samotné struktuře.

11 ZÁVĚR

Cílem této bakalářské práce bylo vypracovat literární rešerši a na základě popsaných evolučních modelů vytvořit program pro stanovení evoluční vzdálenosti řetězců nukleotidů, aminokyselin a kodónů.

V první části jsem se zaměřila na teoretický rozbor obecných témat jako například molekulární evoluce a genetika, základní mutační mechanismy nebo struktura a funkce genu. Tyto informace jsou důležité pro pochopení souvislostí v následujících částech práce.

Dále jsem teoreticky popsala použité evoluční modely, mechanismy jejich výpočtu a vztahy mezi nimi. Konkrétně modely pro výpočet evoluční vzdálenosti DNA sekvencí a proteinových sekvencí v aminokyselinové a kodónové reprezentaci. Dále jsou posány také dva modely pro výpočet evoluční vzdálenosti dle kodónu, které nejsou součásti programového řešení, avšak patří k běžně používaným modelům. Jedna se o metodu o modifikovanou metodu Nei-Gojobori a metodu Li-Wu-Luo.

Praktickou část tvoří samotný program, popis použitých algoritmů a popis grafického uživatelského rozhraní, do kterého jsou implementovány všechny použité algoritmy. Tento prvek umožňuje uživateli snadno volit metody výpočtu a přehledně zobrazit dosažené výsledky.

Nedílnou součástí této práce je analýza dosažených výsledků. Při výběru zdrojových dat jsem byla především limitována délkou porovnávaných sekvencí, jelikož se s délkou sekvencí zvyšuje výpočetní náročnost. Pomocí programu jsem zjistila, že pro vybrané soubory sekvencí, nemá volba modelu v rámci jednoho způsobu hodnocení zásadní vliv. Je ale pravděpodobné, že při zkoumání jiných sekvencí se tento vliv bude významněji projevovat. Také jsem se stručně zmínila o chybách, které mohou při výpočtu nastat.

SEZNAM ZKRATEK

JC69	Jukesův – Cantorův model (1969)
K2P	Kimurův dvouparametrový model
K3ST	Kimurův tříparametrový model
F81	Falsensteinův (Tajimův - Neiův) model
HKY85	Hasegawův – Kishinův – Yanův model
T92	Tamurův model
TN93	Tamurův – Neiův model

SEZNAM PŘÍLOH

Příloha 1	Základní fylogenetické pojmy
Příloha 2	CD s digitální verzí bakalářské práce, zdrojovými soubory programu a zdrojovými GenBank soubory použitých sekvencí

SEZNAM POUŽITÉ LITERATURY

[1] FLEGR, Jaroslav. *Evoluční biologie*. 2., opr. a rozš. vyd. Praha: Academia, 2009, 569 s. ISBN 978-80-200-1767-3.

[2] MACHOLÁN, Miloš. *Základy fylogenetické analýzy*. Vyd. 1. Brno: Masarykova universita, 2014, 289 s. ISBN 978-802-1063-631.

[3] MASATOSHI, Nei a Kumar SUDHIR. *Molecular Evolution and Phylogenetics*. USA: Oxford University Press, 2000. ISBN 0199881227.

[4] HIGGS, Paul G a Teresa K ATTWOOD. *Bioinformatics and molecular evolution*. Malden, MA: Blackwell Pub., 2005, xiii, 365 p. ISBN 14-051-0683-2.

[5] KIMURA, Motoo. *Journal of molecular evolution: A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences*. Japan: National Institute of Genetics, 1980. ISBN 0022-2844. ISSN 0022--2844.

[6] RZHETSKY, Andrej a Masatoshi NEI. *Molecular biology and evolution: Tests of Applicability of Several Substitution Models for DNA Sequence Data*. Chicago: The University of Chicago, 1995. ISBN 0737-4038. ISSN 0737-4038.

[7] TAJIMA, Fumio. *Molecular biology and evolution: Unbiased Estimation of Evolutionary Distance between Nucleotide Sequences*. Chicago: University of Chicago, 1993. ISBN 0737-4038. ISSN 0737-4038.

[8] NEI, Masatoshi a Koichiro TAMURA. *Molecular biology and evolution: Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans and Chimpanzees*. Chicago: University of Chicago, 1993. ISBN 0737-4038. ISSN 0737-4038.

[9] TAMURA, Koichiro. *Molecular biology and evolution: Model Selection in the Estimation of the Number of Nucleotide Substitutions*. Chicago: University of Chicago, 1994. ISBN 0737-4038. ISSN 0737-4038.

[10] ŠKUTKOVÁ, Helena. *Studijní materiály k předmětu Analýza biologických sekvencí*. FEKT VUT, 2014.

[11] NEI, Masatoshi a Takashi GOJOBORI. *Molecular biology and evolution: Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions*. Chicago: University of Chicago, 1986. ISBN 0737-4038. ISSN 0737-4038.

[12] LUO, Chi-Cheng, Chung-I WU a Wen-Hsiung LI. *Molecular biology and evolution: A New Method for Estimating Synonymous and Nonsynonymous Rates of Nucleotide Substitution Considering the Relative Likelihood of Nucleotide and Codon Changes*. Chicago: University of Chicago, 1985. ISBN 0737-4038. ISSN 0737-4038.

[13] YANG, Ziheng. *Computational molecular evolution*. Oxford: Oxford University Press, 2006, xvi, 357 p. ISBN 978-019-8567-028.

[14] *National Center for Biotechnology Information (NCBI)* [online]. [cit. 2015-05-27]. Dostupné z: http://www.ncbi.nlm.nih.gov/

[15] NARUYA, Saitou a Masatoshi NEI. *Molecular biology and evolution: The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees*. Chicago: University of Chicago, 1987. ISBN 0737-4038. ISSN 0737-4038.

[16] SEDGWICK, P. Pearson's correlation coefficient. *BMJ*. 2012, **345**(jul04 1): e4483-e4483. DOI: 10.1136/bmj.e4483. ISSN 1756-1833. Dostupné také z: http://www.bmj.com/cgi/doi/10.1136/bmj.e4483

[17] PATTENGALE, Nicholas D., Eric J. GOTTLIEB a Bernard M.E. MORET. Efficiently Computing the Robinson-Foulds Metric. *Journal of Computational Biology*. 2007, **14**(6): 724-735. DOI: 10.1089/cmb.2007.R012. ISSN 1066-5277. Dostupné také z: http://www.liebertonline.com/doi/abs/10.1089/cmb.2007.R012

[18] STUDIER, James A. a Karl J. KEPPLER. *Molecular biology and evolution: A Note on the Neighbor-Joining Algorithm of Saitou and Neil*. Chicago: University of Chicago, 1988. ISBN 0737-4038. ISSN 0737-4038.

[19] YANG, Ziheng a Rasmus NIELSEN. *Molecular biology and evolution: Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages*. Chicago: University of Chicago, 2002. ISBN 0737-4038. ISSN 0737-4038.

[20] LITSCHMANNOVÁ, Martina**.** *Vybrané kapitoly z pravděpodobnosti*. Ostrava : VŠB – TU Ostrava, Fakulta elektrotechniky a informatiky, 2011.

PŘÍLOHA 1:

Pro snazší pochopení textu je potřeba seznámit se s některými základními pojmy fylogenetiky. Hesla jsou popsána zjednodušeně jen pro účely pochopení práce.

RNA	Z anglického ribonucleic acid - Makromolekula složená z řetězce nukleotidů, obsahujících cukr ribózu. Nejčastěji tvoří jednovlánovou strukturu. Rozlišujeme transverovou RNA (tRNA), mediátorovou RNA (mRNA) a ribozomální RNA (rRNA). Mezi nejdůležitější funkce RNA patří přenos genetické informace při transkripci a translaci.
DNA	Z anglického deoxyribonucleic acid - Makromolekula složená z řetězce nukleotidů, obsahujících cukr deoxyribózu. Nejčastěji tvoří dvoušroubovici, v níž jsou jednotlivé řetězce uspořádány dle komplementarity bází. Je nositelkou genetické informace, kdy díky procesu transkripce do mRNA a následné translace utváří primární strukturu proteinu dle genetického kódu.
NUKLEOTID	Biologické molekuly, skládající se z cukru (ribóza nebo deoxyribóza), fosfátového zbytku a nukleové báze. Názvosloví nukleotidů odpovídá bázi, kterou daný nukleotid obsahuje. Báze adenin, cytosin, guanin a thymin (zkratky A, C, G, T) se vyskytují u DNA. U RNA je thymin nahrazen uracilem (zkratka U). Jednotlivé nukleotidy jsou k sobě vázány fosfodiesterovou vazbou a vytváří řetězec DNA či RNA. Nukleotidy se dále mohou spojovat s dalším řetězcem pomocí vodíkových můstků dle komplementarity bází.
GEN	Specifický úsek DNA, který je exprimován do struktury proteinu. Soubor všech genů tvoří genotyp, který společně s prostředím utváří fenotyp daného organismu.
START KODON	Též iniciační kodon. Kodon, u něhož začíná proces translace. Většinou se jedná o kodon AUG.
STOP KODON	Kodon, u něhož dochází k zastavení translace a tedy i celé proteosyntézy. Jedná se o kodony TAA, TAG, TGA.

PROTEIN	Protein, nebo-li bílkovina, je makromolekula, jejíž primární strukturu tvoří řetězec aminokyselin. Jednotlivé aminokyseliny jsou pospojovány peptidovou vazbou (-NH-CO-). Proteiny tvoří podstatu všech živých organismů.
AMINOKYSELINA	V užším slova smyslu je chápeme jako 23 základních stavebních jednotek proteinů.
TRANSICE	Jedná se o bodovou mutaci, při níž dochází k záměně purinové báze za jinou, taktéž purinovou.
TRANSVERZE	Jedná se o bodovou mutaci, při níž dochází k záměně purinové báze za pyrimidinovou.
ORTOLOGNÍ GENY	Jsou výsledkem speciace původního genu, divergují po vzniku druhu. Čili u všech druhů mají stejnou funkci.