VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

METODY DETEKCE SELEKCE V DNA SEKVENCÍCH
METHODS TO DETECT SELECTION IN DNA SEQUENCES

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Ondřej Procházka**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Helena Škutková, Ph.D.**
SUPERVISOR

BRNO 2016

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní obor **Biomedicínské inženýrství a bioinformatika**
Ústav biomedicínského inženýrství

Student: Bc. Ondřej Procházka ***ID:*** 147470
Ročník: 2 ***Akademický rok:*** 2015/16

NÁZEV TÉMATU:

Metody detekce selekce v DNA sekvencích

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte teoretický úvod do problematiky molekulární evoluce zaměřený zejména na mechanismy šíření a fixace molekulárních znaků. 2) Vypracujte literární rešerši výpočetních metod pro detekci genů a úseků genomů podléhající selekci na molekulární úrovni na základě poměru synonymních a nesynonymních mutací. 3) Vybrané metody realizujte v programovém prostředí Matlab. 4) Sestavte vhodný datový set sekvencí DNA z veřejných databází pro statistické testování vybraných metod. 5) Vytvořte programové rozhraní umožňující zarovnání sekvencí podle ORF, vyhodnocení míry selekce na úrovni genů i celých genomů a detekci úseků podléhající selekci na základě alespoň tří odlišných metod. 6) Na základě výsledků realizovaných metod vypracujte statistické vyhodnocení věrohodnosti molekulární selekce u vytvořeného setu dat.

DOPORUČENÁ LITERATURA:

[1] YANG, Z. a J. P. BIELAWSKI. Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution, 12/1/ 2000, 15(12), 496-503.

[2] HURST, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends in Genetics, 9/1/ 2002, 18(9), 486-487.

Termín zadání: 8.2.2016 ***Termín odevzdání:*** 20.5.2016

Vedoucí práce: Ing. Helena Škutková, Ph.D.
Konzultant diplomové práce:

prof. Ing. Ivo Provazník, Ph.D., *předseda oborové rady*

Abstrakt

Tématem diplomové práce jsou metody detekce selekce v DNA sekvencích. V úvodu se popíše molekulární evoluce. Uvedeme, čím je způsobena a jak se projevuje. Dále budou popsány genové mutace a mechanismy šíření a fixace. Rozebereme si podrobně selekci a obecné matematické vztahy pro selekční tlak. Definujeme pozitivní, negativní a neutrální selekci. Práce se zaměří na výpočetní metody evolučních vzdáleností synonymních a nesynonymních substitucí. Budou popsány tři metody pro výpočet selekčního tlaku – Nei-Gojobori, Li-Wu-Luo a Comeron. Popíšeme veškeré modely matematickými vztahy. Pro statistické vyhodnocení jednotlivých selekcí budou zavedeny statistické testy – využívat se bude z-test. V praktické části budou informace o vytvořeném softwaru, který ze sekvencí z veřejných databází ve formátu GenBank vypočítá selekční tlak a zobrazí oblasti selekce na základě statistického testu. Program bude využit na dva datové sety dvou různých genových kódů. Jejich výsledky budou porovnávány. Zhodnotíme všechny tři metody výpočtu selekčního tlaku a vlivu vstupních parametrů.

Klíčová slova

Mutace, přírodní výběr, selekce, selekční tlak, K_a/K_s poměr, Nei-Gojobori, Li-Wu-Luo, Comeron, synonymní substituce, nesynonymní substituce

Abstract

The topic of semestral thesis is methods to detect selection in DNA sequences. In the begining of the thesis we will describe molecular evolution. It will be written what made the evolution and how the evolution is shown. Moreover there are gen mutations and mechanisms of diffuse and fixation. It will be defined what pozitive, negative and neutral selection is. The thesis is focused on evolution distance of synonymous and nonsynonymous substitution. There will be described three methods – Nei-Gojobori, Li-Wu-Luo and Comeron. All these methods will be described with mathematic formulas. There will be statistic test to decide what kind of selection ti is – there will be used z-test. In the practical part, there will be information about developed software what counts selection pressure from sequences from databazes in format GenBank and it shows parts where selection is. The software will be used for two data sets with two different genetic codes. The result will be discussed. We will discuss results of all three methods of selection pressure and influence of input parametrs.

Key words

Mutation, selection, natural selection, selective pressure, Ka/Ks ratio, Nei-Gojobori, Li-Wu-Luo, Comeron, synonymous substitution, nonsynonymous substitution

PROCHÁZKA, O. Metody detekce selekce v DNA sekvencích. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2016. 78 s. Vedoucí diplomové práce Ing. Helena Škutková Ph.D.

Prohlášení

Prohlašuji, že svoji diplomovou práci na téma Metody detekce selekce v DNA sekvencích jsem vypracoval samostatně pod vedením vedoucího 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 části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009Sb.

V Brně dne 20. 5. 2016 Podpis autora

Poděkování

Děkuji vedoucí diplomové práce Ing. Heleně Škutkové, Ph.D. za účinnou metodickou, pedagogickou a odbornou pomoc při zpracování mé diplomové práce a hlavně za svatou trpělivost a podstatné rady. Dále bych rád poděkoval svému okolí, které mi pomohlo vše dopsat.

V Brně dne 20. 5. 2016 Podpis autora

Obsah

Seznam obrázků vii
Seznam tabulek ix
Úvod 1
1. Molekulární evoluce 2
1.1. Struktura genů 2
1.2. Proteosyntéza 2
1.3. Genetické kódy 3
1.4. Genové mutace 7
1.4.1. Substituce 7
1.4.2. Posunové mutace 10
1.5. Open reading frame 10
1.6. Selekce 11
1.7. Selekční tlak 12
2. Modely výpočtu selekčního tlaku 16
2.1. Nei-Gojobori metoda 16
2.2. Li-Wu-Luo metoda 19
2.3. Comeron metoda 23
2.4. Další metody 25
3. Statistické vyhodnocení výsledků 26
4. Realizace algoritmů 28
4.1. Vstupní parametry k analýze 28
4.2. Předzpracování sekvencí k analýze 29
4.3. Výpočet selekčního tlaku 30
4.4. Zobrazení výsledků 31
5. Data 33
6. Diskuze výsledků 35

7. Závěr 56
Literatura 58
Seznam zkratek 62
Přílohy 63
Příloha č. 1 – Manuál programu 63
Příloha č. 2 – Hodnoty distribuční funkce Φ(x) normované normální náhodné veličiny - výtah 67

Seznam obrázků

Obrázek 1: Nákres proteosyntézy, [5] 2
Obrázek 2: Druhy substituce nukleotidů, [2] 7
Obrázek 3: Druhy substituce, [13] 8
Obrázek 4: Příklad delece 10
Obrázek 5: Příklad ORF 11
Obrázek 6: Příklad pozitivní a negativní selekce, [14] 13
Obrázek 7: Nákres mutace se zvýhodněním sekvence, [16] 14
Obrázek 8: Druhy selekce, [22] 15
Obrázek 9: Oboustranný test, [40] 26
Obrázek 10: Skórovací matice BLOSUM62, [42] 29
Obrázek 11: Savčí cirkulární mtDNA s vyznačením umístění některých genů, [46] 34
Obrázek 12: Výsledky synonymních a nesynonymních distancí mezi *Homo sapiens* a *Bos taurus* – gen BRCA1, velikost okna 100 aa 36
Obrázek 13: Výsledky selekčního tlaku mezi *Homo sapiens* a *Bos taurus* – gen BRCA1, velikost okna 100 aa 36
Obrázek 14: Výsledky selekčního tlaku mezi *Homo sapiens* a *Bos taurus* – gen BRCA1, velikost okna 50 aa 37
Obrázek 15: Výsledky selekčního tlaku mezi *Homo sapiens* a *Bos taurus* – gen BRCA1, velikost okna 200 aa 37
Obrázek 16: Výsledky selekčního tlaku mezi *Homo sapiens* a *Mus musculus* – gen BRCA1, velikost okna 50 aa 38
Obrázek 17: Výsledky selekčního tlaku mezi *Homo sapiens* a *Mus musculus* – gen BRCA1, velikost okna 100 aa 38
Obrázek 18: Výsledky selekčního tlaku mezi *Homo sapiens* a *Mus musculus* – gen BRCA1, velikost okna 200 aa 39
Obrázek 19: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 100 aa 40
Obrázek 20: Výsledky synonymních a nesynonymních distancí mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 100 aa 40
Obrázek 21: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 100 aa, výřez oblasti 230 – 320 aa 41
Obrázek 22: Výsledky synonymních a nesynonymních distancí mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 100 aa, výřez oblasti 230 – 320 aa 41
Obrázek 23: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 50 aa 42

Obrázek 24: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 50 aa, výřez oblasti 230 – 320 aa 42
Obrázek 25: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 200 aa 43
Obrázek 26: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 200 aa, výřez oblasti 230 – 320 aa 43
Obrázek 27: Detekce pozitivní selekce v oknech o velikosti 50 aa u BRCA1 45
Obrázek 28: Detekce pozitivní selekce v oknech o velikosti 100 aa u BRCA1 46
Obrázek 29: Detekce pozitivní selekce v oknech o velikosti 200 aa u BRCA1 46
Obrázek 30: Detekce negativní selekce v oknech o velikosti 50 aa u BRCA1 47
Obrázek 31: Detekce negativní selekce v oknech o velikosti 100 aa u BRCA1 48
Obrázek 32: Detekce negativní selekce v oknech o velikosti 200 aa u BRCA1 48
Obrázek 33: Výsledky selekčního tlaku mezi *Pan paniscus* a *Pan troglodytes* – gen cytochromu b z mtDNA, velikost okna 50 aa 49
Obrázek 34: Výsledky selekčního tlaku mezi *Pan paniscus* a *Pan troglodytes* – gen cytochromu b z mtDNA, velikost okna 100 aa 50
Obrázek 35: Výsledky selekčního tlaku mezi *Pan paniscus* a *Pan troglodytes* – gen cytochromu b z mtDNA, velikost okna 200 aa 50
Obrázek 36: Výsledky synonymních a nesynonymních distancí s velikostí okna 50 aa mezi *Homo sapiens* a *Homo sapiens neanderhalensis* – gen cytochromu b z mtDNA 51
Obrázek 37: Výsledky selekčního tlaku s velikostí okna 50 aa mezi *Homo sapiens* a *Homo sapiens neanderhalensis* – gen cytochromu b z mtDNA 52
Obrázek 38: Výsledky synonymních a nesynonymních distancí s velikostí okna 100 aa mezi *Homo sapiens* a *Homo sapiens neanderhalensis* – gen cytochromu b z mtDNA 52
Obrázek 39: Výsledky selekčního tlaku s velikostí okna 100 aa mezi *Homo sapiens* a *Homo sapiens neanderhalensis* – gen cytochromu b z mtDNA 53
Obrázek 40: Detekce negativní selekce v oknech o velikosti 50 aa u Cytochromu b 53
Obrázek 41: Detekce negativní selekce v oknech o velikosti 100 aa u Cytochromu b 54
Obrázek 42: Detekce negativní selekce v oknech o velikosti 200 aa u Cytochromu b 54
Obrázek 43: Uživatelské prostředí programu 63
Obrázek 44: Prohlížeč souborů se soubory GenBank se sekvencemi 64
Obrázek 45: Výběr genu 64
Obrázek 46: Výběr sekvencí k analýze 64
Obrázek 47: Výběr typu genetického kódu 65
Obrázek 48: Výběr velikosti okna 66
Obrázek 49: Grafické zobrazení progresu analýzy 66

Seznam tabulek

Tabulka 1: seznam aminokyselin dle standardního kódu DNA a jejich kódujících kodonů, [2]3
Tabulka 2: změny u mitochondriálního genetického kódu u obratlovců [37] 4
Tabulka 3: změny u mitochondriálního genetického kódu u kvasinek [37] 4
Tabulka 4: změny u mitochondriálního genetického kódu u obratlovců [37] 5
Tabulka 5: změny u mitochondriálního genetického kódu u bezobratlých [37] 5
Tabulka 6: změny u mitochondriálního genetického kódu u ostnokožců a ploštěnců [37] 5
Tabulka 7: změny u mitochondriálního genetického kódu u sumek [37] 6
Tabulka 8: změny u mitochondriálního genetického kódu u motolic [37] 6
Tabulka 9: Fyzikálně-chemické vlastnosti aminokyselin, [2] 9
Tabulka 10: matice přechodů JC modelu, [13] 19
Tabulka 11: matice přechodů Kimura modelu, [13] 20
Tabulka 12: Fischerův exaktní test, [2] 27
Tabulka 13: IUPAC kódové označení a jejich číselný ekvivalent 30
Tabulka 14: Sekvence vybrané do datového setu genu BRCA1 33
Tabulka 15: Sekvence vybrané do datového setu mtDNA 34
Tabulka 16: Procentuální zastoupení oken s detekcí pozitivní selekce u datového setu BRCA1 [%] 44
Tabulka 17: Procentuální zastoupení oken s detekcí negativní selekce u datového setu BRCA1 [%] 44
Tabulka 18: Procentuální zastoupení oken s detekcí negativní selekce u datového setu mtDNA – Cytochrom b [%] 55
Tabulka 19: Hodnoty distribuční funkce Φ(u), [40] 67

Úvod

Diplomová práce se bude zabývat metodami detekce selekce v DNA sekvencích. Selekce, respektive přírodní výběr je jedna z podstatných informací v evoluci.

V rámci práce prvně shrneme molekulární evoluci, čím je způsobena a jak se projevuje. Obecně popíšeme strukturu genu, přiblížíme problematiku proteosyntézy, která je v práci využívána na zajištění výsledků syntetizovaných aminokyselin, a je připojen i podrobný rozbor jednotlivých genetických kódů, dle kterých dochází k překladu nukleotidů na aminokyseliny. Jsou popsané rozdíly mezi jednotlivými kódy a pro jakou skupinu sekvencí se který používá. Popíšou se genové mutace – druhy substitucí a posunových mutací, a jejich mechanismy šíření a fixace. Základem pro další pochopení jsou informace ohledně synonymních a nesynonymních mutací. Vysvětlí se pojem ORF (open reading frame).

Obecný význam selekce bude rozebrán v jedné z kapitol. Přiblížíme si přírodní výběr, genetický drift a neposlední řadě samotný selekční tlak, který bude popsán dopodrobna. Kapitola bude obsahovat matematické vztahy, které budou vysvětleny a rozebrány. Vše bude popsáno ve vztahu s fixací či eliminací mutací.

Dále se v práci zaměříme na výpočetní modely samotného selekčního tlaku. Pro jejich matematický základ budou popsány výpočty evolučních vzdáleností, respektive modely Jukes-Cantor a Kimura. Po úvodu této části se budou podrobněji popisovat samotné metody pro výpočet selekčního tlaku aproximované zmíněnými modely pro evoluční vzdálenosti. Rozeberou se metody Nei-Gojobori, Li-Wu-Luo a Comeron. Budou popsány celé matematické algoritmy výpočtu těchto metod a přidány jednoduché příklady pro lepší pochopení.

Pro přesnější rozřazení selekce budou zmíněny dva statistické testy, kterými lze určit pozitivní, negativní či neutrální selekci. Podrobněji bude popsán Z-test, ke kterému budou veškeré vzorce již zmíněny u metod výpočtu selekčního tlaku. Zmíníme se i o Fischerově exaktním testu, který ale není v této práci použit.

Na základě všech těchto poznatků bude představen software pro analýzu sekvencí, který je vytvořený v programovém prostředí MATLAB®. Přiblíží se jeho programové prostředí, způsob provedení analýzy.

Před diskuzí výsledků budou popsány datové sety, která byly analyzovány. Následně budou představeny a diskutovány výsledky analýzy již popsanými metodami, které budou porovnávány.

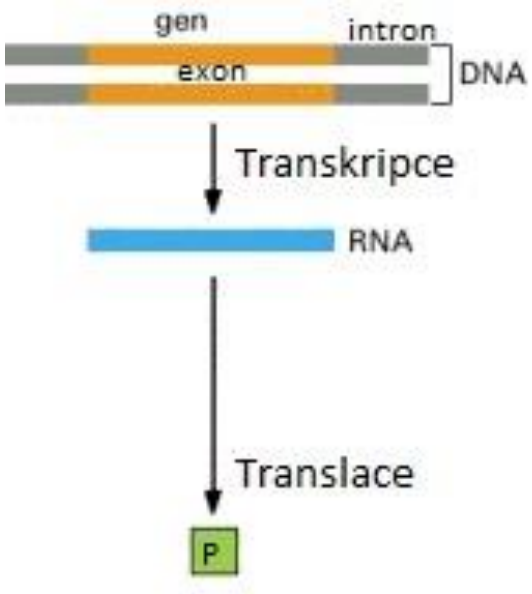
1. Molekulární evoluce

V rámci molekulární evoluce se snažíme zjistit původ a změny v DNA (deoxyribonukleová kyselina) každého genu. Hlavně zkoumáme mechanismy, kterými se nukleotidové sekvence genů mění v závislosti na čase a jejich produkty, především proteiny [1].

1.1. Struktura genů

Základem evolučního procesu je DNA a její změny, mutace v průběhu času. DNA se skládá ze tří stavebních částí – fosfát, cukru 2-deoxy-D-ribózy a jeden ze čtyř dusíkatých bází. Rozlišujeme deriváty purinu, Adenin a Guanin, a deriváty pyrimidinu, Cytosin a Tymin. Dále báze dělíme na základě párování mezi sebou dle počtu vodíkových můstků, 2 můstky pojí Adenin a Tymin a 3 můstky se nachází mezi Cytosinem a Guaninem. Pro přepis z DNA na sekvenci aminokyselin je důležité vzpomenout další komponent – RNA (ribonukleová kyselina). Oproti DNA se skládá z cukru D-ribózy a také z dusíkatých bází, s výjimkou Tyminu, který je u RNA zaměněn za Uracil [2], [3].

1.2. Proteosyntéza



Obrázek 1: Nákres proteosyntézy, [5]

Geny rozlišujeme na skupinu protein-kódující a RNA-kódující. Pro evoluci jsou důležité geny kódující proteiny. Na DNA vláknu určujeme části označující za exony a introny. Exon je oblast DNA kódu, který je prvotně překódován, transkripcí, na mRNA, mediátorovou RNA, a

následně přepsán kód v RNA do sekvencí aminokyselin proteinů. Tento proces se nazývá translace – viz Obrázek 1 [5].

1.3. Genetické kódy

Standardní genetický kód je univerzální pro prokaryotické i eukaryotické buňky. Ze čtyř nukleotidů (adeninu, cytosinu, guaninu a uracilu u RNA, resp. tyminu u DNA) lze vytvořit kombinací až 64 kodonů ($4^3 = 64$). Samotných aminokyselin se kóduje 20 a dále rozeznáváme terminační kodony, takzvané „stop kodony" – kombinace nukleotidů TAA, TAG, TGA. Speciální označení nese i kodon ATG, který je označován za iniciační triplet. Jelikož aminokyselin není takové množství jako kombinací nukleotidů, kóduje více kodonů stejnou aminokyselinu – viz Tabulka 1[2], [5].

Tabulka 1: seznam aminokyselin dle standardního kódu DNA a jejich kódujících kodonů, [2]

Kodón	Aminokys.	Zk.		Kodón	Aminokys.	Zk.		Kodón	Aminokys.	Zk.		Kodón	Aminokys.	Zk.
TTT	Fenylalanin	Phe (F)		TCT	Serin	Ser (S)		TAT	Tyrosin	Tyr (Y)		TGT	Cystein	Cys (C)
TTC	Fenylalanin	Phe (F)		TCC	Serin	Ser (S)		TAC	Tyrosin	Tyr (Y)		TGC	Cystein	Cys (C)
TTA	Leucin	Leu (L)		TCA	Serin	Ser (S)		TAA	Terminační	*		TGA	Terminační	*
TTG	Leucin	Leu (L)		TCG	Serin	Ser (S)		TAG	Terminační	*		TGG	Tryptofan	Trp (W)
CTT	Leucin	Leu (L)		CCT	Prolin	Pro (P)		CAT	Histidin	His (H)		CGT	Arginin	Arg (R)
CTC	Leucin	Leu (L)		CCC	Prolin	Pro (P)		CAC	Histidin	His (H)		CGC	Arginin	Arg (R)
CTA	Leucin	Leu (L)		CCA	Prolin	Pro (P)		CAA	Glutamin	Gln (Q)		CGA	Arginin	Arg (R)
CTG	Leucin	Leu (L)		CCG	Prolin	Pro (P)		CAG	Glutamin	Gln (Q)		CGG	Arginin	Arg (R)
ATT	Isoleucin	Ile (I)		ACT	Threonin	Thr (T)		AAT	Kys. asparagová	Asn (N)		AGT	Serin	Ser (S)
ATC	Isoleucin	Ile (I)		ACC	Threonin	Thr (T)		AAC	Kys. asparagová	Asn (N)		AGC	Serin	Ser (S)
ATA	Isoleucin	Ile (I)		ACA	Threonin	Thr (T)		AAA	Lysin	Lys (K)		AGA	Arginin	Arg (R)
ATG	Methionin	Met (M)		ACG	Threonin	Thr (T)		AAG	Lysin	Lys (K)		AGG	Arginin	Arg (R)
GTT	Valin	Val (V)		GCT	Alanin	Ala (A)		GAT	Asparagin	Asp (D)		GGT	Glycin	Gly (G)
GTC	Valin	Val (V)		GCC	Alanin	Ala (A)		GAC	Asparagin	Asp (D)		GGC	Glycin	Gly (G)
GTA	Valin	Val (V)		GCA	Alanin	Ala (A)		GAA	Kys. glutamová	Glu (E)		GGA	Glycin	Gly (G)
GTG	Valin	Val (V)		GCG	Alanin	Ala (A)		GAG	Kys. glutamová	Glu (E)		GGG	Glycin	Gly (G)

Mimo standardní genetický kód rozlišujeme dalších 18 typů kódů s různými odlišnostmi v přepisu genetického kódu na aminokyselinový zápis. Jedná se o translaci u jiných typů organismů či jiného typu DNA – nejčastěji při zdrojovém genetickém kódu z mitochondrií.

K rozdílnému překladu z genetického kódu do aminokyselin dochází u mitochondriálního genetického kódu obratlovců – viz Tabulka 2. Dále se při tomto genetickém kódu zjistili alternativní inicializační kodony – ATA u tura, ATA a ATT u člověka a navíc u myši ATC [24][25].

Tabulka 2: změny u mitochondriálního genetického kódu u obratlovců [37]

Kodón	Standardní genetický kód	Mitochondriální genetický kód u obratlovců
AGA	Arginin	Terminační kodón
AGG	Arginin	Terminační kodón
ATA	Isoleucin	Methionin
TGA	Terminační kodón	Tryptofan

U vybraných druhů kvasinek - *Saccharomyces cerevisiae, Candida glabrata, Hansenula saturnus*, a *Kluyveromyces thermotolerans*) – dochází ke změnám v translaci na aminokyseliny – viz Tabulka 3 [23].

Tabulka 3: změny u mitochondriálního genetického kódu u kvasinek [37]

Kodón	Standardní genetický kód	Mitochondriální genetický kód u kvasinek
ATA	Isoleucin	Threonin
CTT	Leucin	Threonin
CTC	Leucin	Threonin
CTA	Leucin	Threonin
CTG	Leucin	Threonin
TGA	Terminační kodón	Tryptofan
CGA	Arginin	chybí
CGC	Arginin	chybí

Mitochondriální DNA kód plísní, prvoků, žahavců a rodu bakterií *Mycoplasma* se od standardního kódu liší v překladu na aminokyseliny pouze v jednom případě – TGA triplet se překládá na Tryptofan místo terminačního kodonu. Bylo zjištěno, že u této skupiny se vyskytuje mnoho různých alternativních inicializačních kodonů - ATT, ATA, ATG, ATC, CTG, GTG, GTA, TTA a TTG [26].

U mitochondriální DNA bezobratlých je nejčastějším zástupcem Octomilka, u které se triplet AGG nenachází. Dále se u tohoto genetického kódu považují za alternativní začátky překladu ATA, ATT, ATC a TTG. Změny oproti standardnímu kódu se nacházejí u stejných kodonů jako

u mitochondriálního genetického kódu obratlovců, ale Arginin u standardního kódu se nově překládá na Serin. Změny jsou vyznačeny – viz Tabulka 5 [24] [25]

Tabulka 5: změny u mitochondriálního genetického kódu u bezobratlých [37]

Kodón	Standardní genetický kód	Mitochondriální genetický kód u bezobratlých
AGA	Arginin	Serin
AGG	Arginin	Serin
ATA	Isoleucin	Methionin
TGA	Terminační kodón	Tryptofan

Další odlišnosti v translaci se vyskytují u kódů DNA nálevníků, rodu zelených řas z třídy *Dasycladophyceae* a řádu jednobuněčných eukaryotických organismů *Diplomonadida*. Jejich změna je u terminačních kodonů – viz Tabulka 4 [27].

Tabulka 4: změny u mitochondriálního genetického kódu u obratlovců [37]

Kodón	Standardní genetický kód	Mitochondriální genetický kód u nálevníků, *Dasycladophyceae* a *Hexamita*
TAA	Terminační kodón	Glutamin
TAG	Terminační kodón	Glutamin

Mitochondriální genetický kód ostnokožců a ploštěnců se liší od standardního genetického kódu v překladu na aminokyseliny ve 4 případech – viz Tabulka 6. Existuje ještě alternativní překlad pro mitochondriální DNA ploštěnců pro druhy *Radopholus similis* a *Radopholus arabocoffeae*, kde triplet TAA se překládá na Tyrozin oproti terminačnímu kodonu u standardního kódu [29], [33].

Tabulka 6: změny u mitochondriálního genetického kódu u ostnokožců a ploštěnců [37]

Kodón	Standardní genetický kód	Mitochondriální genetický kód u ostnokožců a ploštěnců
AAA	Lysin	Kys. asparagová
AGA	Arginin	Serin
AGG	Arginin	Serin
TGA	Terminační kodón	Tryptofan

V DNA specifického druhu kmenu nálevníci a podtřídy *Hypotrichs – Euplotidae* se mění oproti standardnímu kódu pouze jeden terminační kodon TGA na kodon kódující aminokyselinu Cystein. U bakteriálních a rostlinných plastidů se oproti standardnímu kódu nemění nic. Jsou zjištěné další možné inicializační triplety – GTG, TTG, CTG, NTG (N značí jakýkoli nukleotid), ATT a TGA. Sekvence kvasinek s alternativní jadernou DNA se od standardního kódu mění u tripletu CTG, kdy dochází k translaci na Serin místo Leucinu. Alternativním inicializačním kodonem v tomto případě je CAG u organismu kandida bělostná [27], [30],[31].

U mitochondriální DNA sumek se vyskytují odlišnosti v přepisu na aminokyseliny obdobně jako u bezobratlých – viz Tabulka 7. ATA, GTG, TTG a ATT jsou dalšími alternativními inicializujícími kodony [32].

Tabulka 7: změny u mitochondriálního genetického kódu u sumek [37]

Kodón	Standardní genetický kód	Mitochondriální genetický kód u sumek
AGA	Arginin	Glycin
AGG	Arginin	Glycin
ATA	Isoleucin	Methionin
TGA	Terminační kodón	Tryptofan

Další kategorií rozdílného překladu na aminokyseliny je mitochondriální DNA zelených řas třídy zelenivky. Zde se nachází jediná změna oproti standardu. Dochází k překladu tripletu TAG na Leucin [34].

Velkou skupinou změn v překladu tripletů oproti standardnímu kódu je mitochondriální DNA motolic. Různost je zobrazena v Tabulka 8. Jedná se o stejné změny jako u bezobratlých, jelikož jsou jednou třídou z nich, ale je zde rozšíření o změnu u tripletu AAA [35].

Tabulka 8: změny u mitochondriálního genetického kódu u motolic [37]

Kodón	Standardní genetický kód	Mitochondriální genetický kód u motolic
AGA	Arginin	Serin
AGG	Arginin	Serin
ATA	Isoleucin	Methionin
AAA	Lysin	Kys. asparagová
TGA	Terminační kodón	Tryptofan

Menší rozdíly v translaci nukleotidů na aminokyseliny se vyskytují také u řetízovek, v jejich mitochondriální DNA. Dochází zde k překladu tripletu TCA, standardně vytváří Serin, na terminační kodon a ke změně u TAG kodonu, který u tohoto genetického kódu není stop

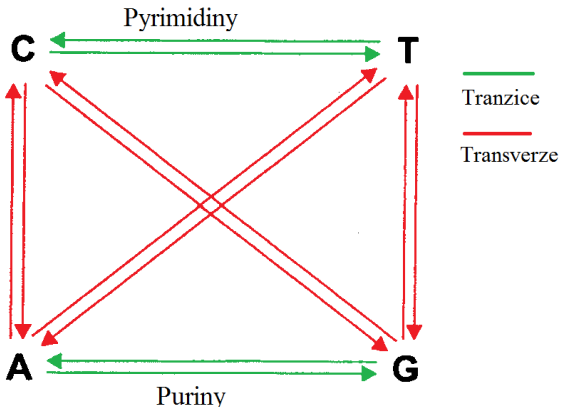
kodonem, jak je standardně, ale dochází k překladu na Leucin. Dalšími organismy, u kterých jsou rozdíly v překladu DNA, jsou z řádu *Thraustochytriales*, třídy *Labyrinthulomycetes* (česky: labyrintuly, čili „vodní hlenky“). Jejich odlišnost od kódu bakterií je založena na změně TTA kodonu v terminační kodon a inicializační triplety jsou ATT, ATG a GTG [36] [37].

1.4. Genové mutace

Tyto mutace probíhají na úrovni vláken DNA. Jedná se o soubor bodových mutací. Vznikají v exonech i intronech. Většinově se objevuje více mutací v oblastech intronů než exonů. Rozlišujeme tři základní typy změny DNA – substituce, inverse, delece. V některých materiálech se objevuje i inverze [2], [3], [5], [7].

1.4.1. Substituce

Jeden nebo více DNA bází je zaměněno za jinou DNA bázi. Substituce nukleotidů se dělí dle toho, zdali se báze zamění za druhou bázi stejného typu – purin za purin (A –> G a G –> A) a pyrimidin za pyrimidin (C –> T a T –> C). Takto proběhlá substituce se nazývá tranzice. Druhým druhem je transverze, kdy se báze mění za báze z druhé skupiny – purin na pyrimidin a naopak (A –> C, A –> T, G –> C, G –> T, C –> A, C –> G, T –> A a T –> G). Dle četnosti je frekventovanější tranzice. Názorně jsou obě druhy substituce zobrazeny na Obrázek 2 [2], [3], [6].

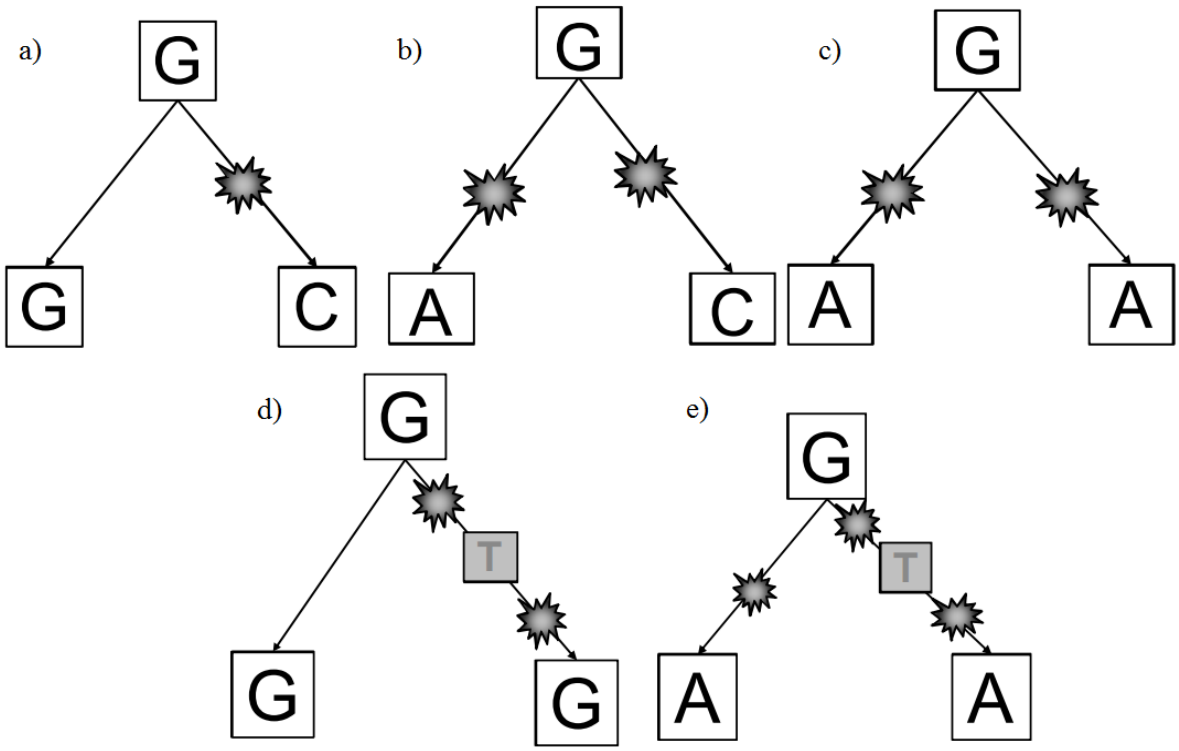


Obrázek 2: Druhy substituce nukleotidů, [2]

Během evoluce genu společného předka jsou možné různé druhy substituce na potomky. Rozdělujeme je na:

- Jednoduché (Obrázek 3, a)) – jedna substituce u sekvence
- Vícenásobná – postupné substituce na jednom místě, výsledkem je změna nukleotidu
- Souběžná (Obrázek 3, b)) – dojde o substituci u homologických DNA sekvencích na stejném místě na různé nukleotidy

- Paralelní (Obrázek 3, c)) – stejná substituce stejného nukleotidu na stejném místě homologických DNA
- Zpětná (Obrázek 3, d)) – návrat uskutečněné substituce na původní nukleotid
- Sbíhavá (Obrázek 3, e)) – stejná substituce různých nukleotidů ve stejném místě homologických DNA



Obrázek 3: Druhy substituce, [13]

Podle ovlivnění výsledného produktu proteosyntézy rozdělujeme substituční změny DNA na synonymní, neměnící smysl, a nesynonymní mutace, měnící smysl. Může dojít ještě k nesmyslné mutaci, kdy se mutací stane triplet stop kodonem a dojde k předčasnému ukončení syntézy peptidu a výsledkem je zcela nefunkční produkt.

Synonymní mutace odpovídá mutaci kodonu, kterou se nezmění výsledný protein tripletu. V tomto případě dochází k mutaci třetího nukleotidu v kodonu. Jelikož více trojic vytvoří stejnou aminokyselinu (viz Tabulka 1), je tato mutace „tichá“, neovlivňuje výsledný protein. Příkladem může být změna originálního kodonu pro Izoleucin ATT na ATC, respektive ATA, kdy při proteosyntéze dojde ke stejnému výsledku. Z hlediska přirozeného výběru jsou neviditelné, neutrální. Pro organismus může mít ale význam, z jakého tripletu je aminokyselina kódována. Ovlivňuje to například sekundární strukturu syntetizované RNA, stabilitu i intenzitu a rychlost translace [1], [2], [3], [7], [8].

Mutací nesynonymní se označují změny v kodonech, které pozmění kódovanou aminokyselinu. Nejčastěji se jedná o mutace na první či druhé pozici v kodonech. Příkladem může být změna druhého nukleotidu kodonu Histidinu CAU na CGU kódující Arginin. Při změně aminokyseliny se určuje, zdali se jedná o změnu na aminokyselinu s podobnými vlastnosti (tzv. konzervativní záměna) nebo s rozdílnými vlastnostmi (tzv. nekonzervativní záměna). Toto označení vychází z fyzikálně-chemických vlastností samotných proteinogenních aminokyselin – viz Tabulka 9 [2], [8].

Tabulka 9: Fyzikálně-chemické vlastnosti aminokyselin, [2]

Kódy		Jméno	Fyzikálně-chemické vlastnosti při pH 7
A	Ala	Alanin	nepolárni, hydrofobni
C	Cys	Cystein	polárni
D	Asp	Kys. Asparagová	polárni, hydrofilni, kyselina
E	Glu	Kys. Glutaminová	polárni, hydrofilni, kyselina
F	Phe	Fenylalanin	nepolárni, hydrofobni
G	Gly	Glycin	nepolárni
H	His	Histidin	polárni, hydrofilni, esenciálni
I	Ile	Isoleucin	nepolárni, hydrofobni
K	Lys	Lysin	polárni, hydrofilni, esenciálni
L	Leu	Leucin	nepolárni, hydrofobni
M	Met	Methionin	nepolárni, hydrofobni
N	Asn	Asparagin	polárni, hydrofilni, neutrálni
P	Pro	Prolin	nepolárni
Q	Gln	Glutamin	polárni, hydrofilni, neutrálni
R	Arg	Arginin	polárni, hydrofilni, esenciálni
S	Ser	Serin	polárni
T	Thr	Threonin	polárni
V	Val	Valin	nepolárni, hydrofobni
W	Trp	Tryptofan	nepolárni
Y	Tyr	Tyrosin	polárni

1.4.2. Posunové mutace

Mutační změny způsobují i změny počtu nukleotidů v DNA řetězci. K tomu dochází, pokud proběhne delece nebo inzerce. Pokud změna nukleotidů je o jiný než 3n počet, tak v důsledku takové změny se posune čtecí rámec. Následně se celý řetězec, upravený pouze o 1 či 2 nukleotidy, překládá v naprosto jiný produkt. Může dojít i k vzniku terminačního kodonu dříve než v originální sekvenci – viz Obrázek 4. V oblastech intronů se objevují delece a inzerce často. K těmto mutačním změnám dochází kvůli chybám v DNA replikaci. Dlouhé úseky inzerce se již zařazují mezi transpozice, kdy při křížení DNA se zařadí jedna část cizí DNA do druhé [2], [5], [6], [7].

Příklad delece jednoho nukleotidu

původní DNA	CCG	TAT	ACG	TGC	AAT	CGA	TAC
mRNA	GGC	AUA	UGC	ACG	UUA	GCU	AUG
protein	Gly	Ile	Cys	Thr	Leu	Ala	Met

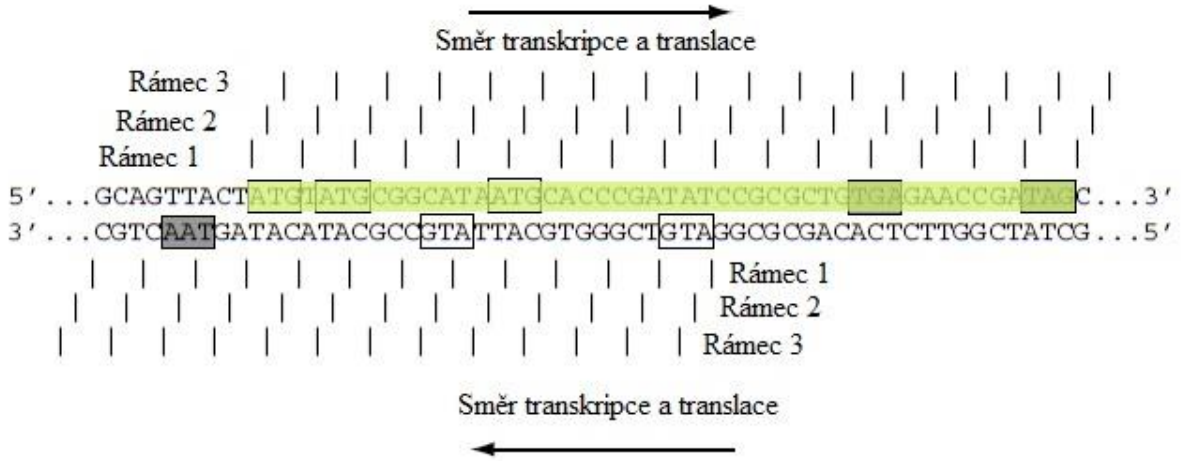
zmutovaná DNA	CCG	TAT	ACG	GCA	ATC	GAT	AC
mRNA	GGC	AUA	UGC	CGU	UAG	CUA	UG
protein	Gly	Ile	Cys	Arg	**Ter.**	Leu	-

Obrázek 4: Příklad delece

1.5. Open reading frame

Označení ORF odpovídá zkratce z anglického open reading frame – což značí část sekvence, kde se nacházejí triplety kódující protein. Jejich velikost je závislá na počátečním, inicializačním kodonu a posléze ukončovacím stop kodonu. U každého genetického kódu jsou tyto triplety různé, což bylo zmíněno již výše. V rámci jedné části sekvence se mohou ORF oblasti nacházet v různých místech. Rozeznáváme 6 možných kombinací k nalezení ORF. Lze hledat od 1., 2. či 3. nukleotidu v DNA sekvenci nebo u reverzní DNA sekvence. Jelikož u genů kódujících specifický protein jsou již tyto části nalezeny a popsány, označují se ve veřejných databázích jako CDS oblasti. V informaci o sekvenci jsou tyto CDS oblasti zaznamenány. V sekvencích se nachází mnoho ORF a odpovídají v podstatě exonům – viz Obrázek 5. V příkladu se nachází inicializační kodon ATG na mnoha místech v sekvenci. Terminační kodony TGA, AAT potenciálně ukončují transkripci. Využito je pouze takového ORF, které

kóduje smysluplný překlad. V případě příkladu by se využilo rámce 1 ve směru sekvence 5‘-3’konce [39].



Obrázek 5: Příklad ORF

1.6. Selekce

Evoluce na molekulární úrovni je způsobena výběrem, selekcí. Přírodní výběr dle Darwina znamená, že se rozmnožují nejlépe přizpůsobení jedincí, kteří jsou životaschopnější, průbojnější, silnější, rychlejší, seženou více potravy, uniknou dříve predátorovi, případně jsou atraktivnější pro samičku. Do dalších generací se tedy dostává jen genetická informace od těch nejschopnějších jedinců a tím dochází k vývoji. Dalším faktorem ovlivňující vývoj je genetický drift a selekční tlak při náhodně způsobené mutaci genů.

Za drift se bere náhodný výběr genetického materiálu v dané populaci. Rozumí se zde, že dochází k náhodné fluktuaci v četnosti alel postupně se projevující zvyšováním této četnosti. V rámci dědičnosti může docházet k potlačení jednoho ze dvou typů genetického materiálu a preference druhého. Obvykle je označován za polymorfizmus. O náhodný genetický posun lze mluvit jen za předpokladu, že absolutní hodnota selekce u příslušné alely je nižší než $1/N$, resp. pro diploidní organismy $1/2N$, kde N je počet genů. Neutrální selekce dává rovnostářskou povahu propagaci mutantů v populaci. Drift není vždy považován za pozitivní selekci. Může docházet i k driftu, který vede ke kvalitativnímu zhoršení jedince. Drift dopomáhá k fixaci nebo eliminaci změn v populaci.

Pravděpodobnost fixace mutace jednoho genu je vyjádřena vztahem (1.1) pro diploidní organismy.

$$u = \frac{1}{2N} \tag{1.1}$$

Z neutrální teorie je rychlost evoluční substituce rovna počtu nových mutantů ($2N\mu$, kde μ označuje poměr mutací na gen v generaci) násobeného pravděpodobností jejích fixace – u. Z toho vyplývá vztah pro rychlost evoluční substituce – λ – viz rovnice (1.2). Po dosazení se rychlost evoluční substituce rovná proměnné μ - viz vztah (1.3), tedy je nezávislá na jiných faktorech, jakož jsou velikost populace či čas [1], [16].

$$\lambda = 2N\mu u \tag{1.2}$$

$$\lambda = 2N\mu \frac{1}{2N} = \mu \tag{1.3}$$

1.7. Selekční tlak

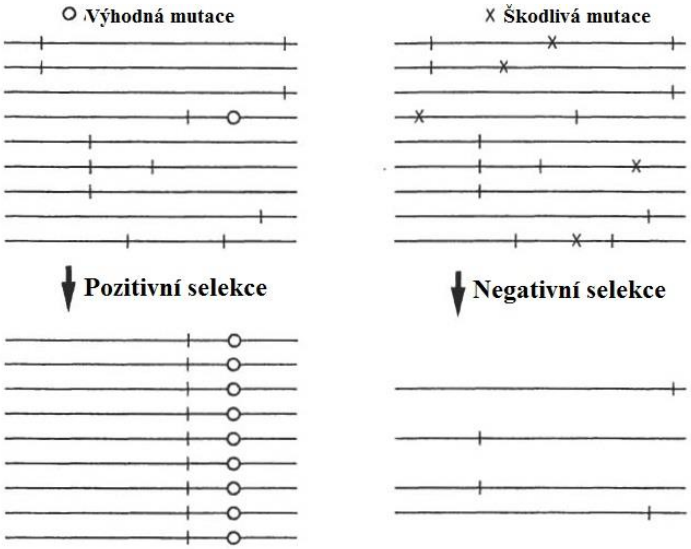
Selekční tlak je poměr evolučních vzdáleností synonymních substitucí, kdy nukleotidová substituce nemění význam a výslednou kódovanou aminokyselinu, a nesynonymních substitucí, kdy změna nukleotidu způsobí výslednou translaci na novou aminokyselinu. Pokud alela zvýhodňuje jedince, bude frekvence této alely významně vyšší [14].

Selekční tlak ω se stanovuje poměrem odhadu dle vzorce (1.4), kde d_N odpovídá nesynonymní nukleotidové diferenci v sekvencích a d_S odpovídá synonymní nukleotidové diferenci v sekvencích [8].

$$\omega = \frac{d_N}{d_S} \tag{1.4}$$

Z definice synonymních substitucí plyne, že vždy dochází k neutrální mutaci. Při aplikaci neutrální teorie, respektive vzorce (1.3), dostaneme, že počet mutací fixovaných na generaci ze synonymních mutací (resp. d_S) můžeme zaměnit za μ, poměr mutací na gen v generaci. Pokud se pro nesynonymní substituce uvažuje, že dochází také k neutrálním mutacím, můžeme obdobně počet mutací fixovaných na generaci z nesynonymních mutací (resp. d_N) zaměnit za μ dle neutrální teorie. Z toho vyplývá vzorec (1.5), tedy že hodnota selekčního tlaku je rovna 1, a je tímto dokázáno, že při této hodnotě dochází k neutrálním mutacím. V takovém případě dochází k tomu, že pozitivní selekce ruší vliv purifikace negativní selekce, a substituce jsou způsobeny přírodním výběrem [4], [16], [18], [20].

$$\omega = \frac{d_N}{d_S} = \frac{\mu}{\mu} = 1 \tag{1.5}$$



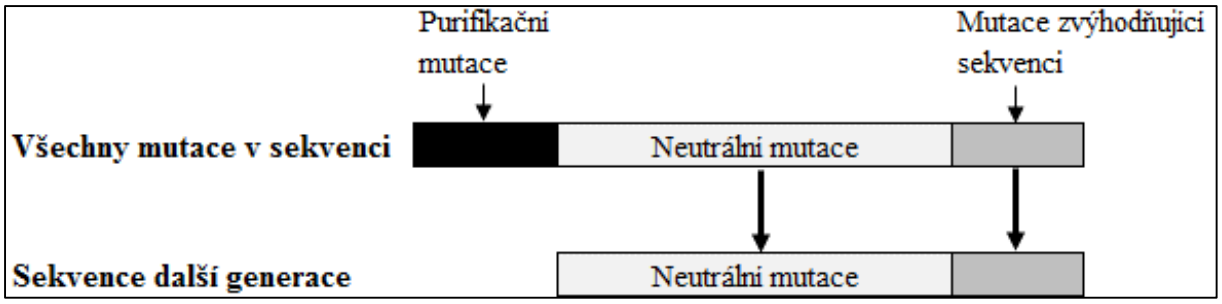
Obrázek 6: Příklad pozitivní a negativní selekce, [14]

Pokud při nesynonymních mutacích dochází ke škodlivým mutacím, respektive eliminaci sekvencí, jedná se o purifikaci – viz Obrázek 6. Selekční tlak v takovém případě je roven hodnotě pod 1 a jedná se o negativní selekci. Hodnotu d_N v tomto případě lze vypočítat dle vzorce (1.6), kde část nesynonymních mutací je neutrálních (označeno jako *f*), která je fixována s poměrem mutací na gen v generaci μ, a zbytek zapříčiňuje purifikaci (část *1-f*). Část nesynonymních substitucí purifikujících sekvence *f* je menší než 1, tedy výsledný selekční tlak je menší než 1 – viz vzorec (1.7) [16], [18].

$$d_N = f\mu + (1-f)0 = f\mu \tag{1.6}$$

$$\omega = \frac{d_N}{d_S} = \frac{f\mu}{\mu} = f < 1 \tag{1.7}$$

Nesynonymní mutace způsobují purifikaci, neboli zánik sekvence, či můžou zvýhodnit jedince a tato nesynonymní mutace může být v populaci zachována, fixována. Obecný výpočet nesynonymní mutace d_N v tomto případě rozdělíme na několik částí – viz (1.8). Názorně je tento případ zobrazen na Obrázek 7. První část způsobuje eliminaci sekvencí opět s proporcionální hodnotou *1-f* bez fixace v generacích. Další část *f* z předchozího případu při negativní selekci nyní rozdělíme na jednu část o velikosti $f\theta$, která způsobuje výhodu sekvenci, a na druhou část o velikosti $f(1-\theta)$, která je neutrální. Při neutrální mutaci uvažujeme s rychlostí fixace μ. Při zvýhodnění sekvence uvažujeme fixaci rovnu hodnotě $2N\mu s$, obdobného vzorci (1.2), kde je pravděpodobnost fixace mutace jednoho genu *u* zaměněna za pravděpodobnost vyjádřenou *s* [14],[16], [18].



Obrázek 7: Nákres mutace se zvýhodněním sekvence, [16]

$$d_N = (1-f)0 + f(1-\theta)\mu + f\theta 2N\mu s \tag{1.8}$$

Z předešlého vyplývá, že výsledek selekčního tlaku při nesynonymních mutacích, které způsobují výhodu sekvenci s pravděpodobností θ, bude odpovídat vzorci (1.9).

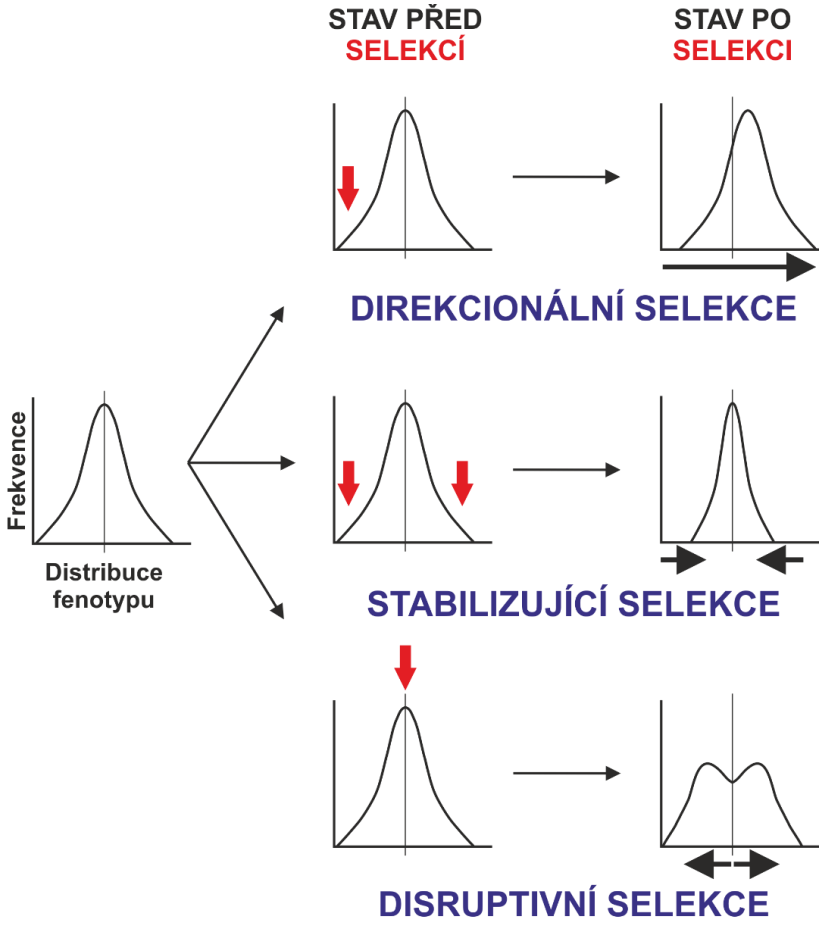
$$\omega = \frac{d_N}{d_S} = \frac{f(1-\theta)\mu + f\theta 2N\mu s}{\mu} = f(1-\theta) + f\theta 2Ns \tag{1.9}$$

Aby selekční tlak byl roven $\omega > 1$, musí být proporcionální část nesynonymních substitucí se zvýhodněním sekvence θ dostatečně vysoká. Ze vzorce (1.9) plyne, že θ musí nabývat hodnoty:

$$\theta > \frac{1-f}{f}\frac{1}{(2N-s)} \tag{1.10}$$

V tomto případě se jedná o pozitivní selekci, která působí ve prospěch adaptivních změn – viz Obrázek 6 [14], [18].

Pozitivní selekci můžeme rozdělit ještě na dva různé typy – přímá selekce (z angl. directional selection) a disruptivní či balancující (z angl. balancing selection) – viz Obrázek 8. Při přímé selekci je zvýhodněn pouze jediný typ alely z populace, buďto průměrného fenotypu, kdy mluvíme přesněji o stabilizující přímé selekci, nebo určitého extrému, kdy se tento jev označuje za usměrňující pozitivní selekci. Přímá selekce tedy upřednostňuje jeden typ alel, který fixuje, a potlačuje genetickou variabilitu. Druhý typ pozitivní selekce, balancující, zvýhodňuje více alel zároveň, které mohou v populaci koexistovat – takový případ bývá označován za heterogenní populaci. Je to velice závislé na frekvenci výskytu alel v populaci a na okolním podmínkách – prostředí aj. [16], [19], [20].



Obrázek 8: Druhy selekce, [22]

2. Modely výpočtu selekčního tlaku

Pro popis evoluční divergence DNA sekvencí se počítají synonymní a nesynonymní mutace zvlášť. Přepočtem na celkový počet všech synonymní či nesynonymní míst se vypočtou hodnoty poměru synonymních mutací na synonymní místa v sekvenci - r_S (někde udávané K_s) a poměru nesynonymních mutací na nesynonymní místa v sekvenci - r_N (respektive K_a). Tyto hodnoty jsou vztažené na generaci či rok. Jelikož čas divergence mezi DNA sekvencemi není vždy znám, využívá se přepočet na diferenciální hodnoty d_S a d_N – výpočet viz rovnice (2.1). Jelikož její výpočet je vztažen k času *t*, zavedly se výpočty evolučních distancí, které jsou popsány v dalších kapitolách [2], [8], [9].

$$d_S = 2r_S t;\ d_N = 2r_N t \tag{2.1}$$

K popisu podobnosti dvou sekvencí se využívají kvalitativní modely. Po zarovnání sekvencí jsou zhodnoceny počty mutací. Proporcionální vzdálenost (zkráceně p-distance) je relativní podíl rozdílných míst, kde dochází k bodovým mutacím, k délce sekvencí. Na základě přechodů jsou vytvořeny matice mezi-stavů. Jednoduchým evolučním modelem, po aplikaci pravděpodobnostního rozložení na proporcionální vzdálenost, je model Jukes-Cantor (JC). Dalším modelem zohledňující frekvence tranzice a transverze je Kimurův model. Komplexnějším evolučním modelem je Tamurův model, ale ten nebude v této práci využit [13].

V této kapitole jsou rozvedeny metody Nei-Gojobori, Li-Wu-Luo a Comeron. Mimo tyto metody jsou známé další, které nejsou předmětem zkoumání v této práci. Poznámka o nich bude na konci kapitoly.

2.1. Nei-Gojobori metoda

Jedná se o metodu založenou na evoluční cestě jednotlivých DNA sekvencí. Počítá s celou možnou evoluční cestou mezi každým párem kodonu mezi dvěma DNA sekvencemi. Předpokládá pravděpodobnost všech nukleotidových změn.

Prvním krokem propočtu je nutné vypočíst přiměřenou možnost synonymní (tiché) změny – označované s - viz rovnice (2.2), tedy, že se nezmění smysl překladu na aminokyselinu. Jelikož každý kodon je složen ze tří nukleotidů, je nutné vypočítat každou možnost v rámci pozice nukleotidu – f_i. Každá pozice může být změněna na tři další nukleotidy, tedy počet synonymních změn je dělen počtem nukleotidů, které nezpůsobují změnu na terminační kodon.

$$s = \sum_{i=1}^{3} f_i, i = 1, 2, 3. \tag{2.2}$$

Př.: výpočet s hodnoty při tripletu CGA (Arg): $s = \frac{1}{2} + \frac{0}{3} + \frac{3}{3} = 1{,}5$, tedy na první pozici se vyskytuje jedna synonymní změna, z nukleotidu C na A – vznikne triplet AGA (Arg), a změna z C na T způsobí terminační kodon. Na druhé pozici nevznikne záměnou nukleotidu nikdy opět Arg aminokyselina. Při záměně třetího nukleotidu vznikne naopak vždy aminokyselina Arg. Vše se řeší pro standardní genetický kód

Výpočet nesynonymních změn jednoho kodonu je nyní pouze odečtením synonymních změn od všech možných – viz rovnice (2.3). Celková hodnota synonymních míst (*S*) pro celou sekvenci je suma *s* každého kodonu – viz rovnice (2.4), a následné vypočtení nesynonymních míst taktéž pro celou sekvenci je vypočteno rozdílem – viz rovnice (2.5). Při aproximaci pro porovnání dvou sekvencí je *S* vyděleno počtem sekvencí, tedy dvěma, jelikož se porovnávají vždy dvě sekvence vůči sobě.

$$n = 3 - s. \tag{2.3}$$

$$S = \sum_{j=1}^{C*3} s_{ja}, \tag{2.4}$$

kde *C* je počet kodonů v sekvenci, a je počet sekvence.

$$N = 3C - (\frac{S}{a}). \tag{2.5}$$

Následující část výpočtu se zaobírá změnou jednoho kodonu první DNA sekvence na kodon na stejném místě v druhé DNA. Počítají se opět rozdílné mutace synonymní a nesynonymní. U dvojice kodonů je potřeba zjistit, na kolika místech se stala mutační změna. Počet mutačních cest, kterými mohl kodon projít ke změně na kodon druhé sekvence DNA, je vyjádřen vzorcem (2.6). Z výsledných cest se sečte počet synonymních a nesynonymních mutací a podělí se počtem mutačních cest *P(k)*. Pokud se při postupných mutacích v nějaké cestě vyskytne terminační kodon, tato cesta se nepočítá do výsledného výpočtu. Jmenovatel se o počet cest s terminačními kodony zmenší. Tento zlomek vyjadřuje s_d pro synonymní a n_d pro nesynonymní mutace jednoho kodonu. Výsledné S_d a N_d je suma parciálních propočtů s_d a n_d pro jednotlivé kodony DNA sekvencí.

$$P(k) = k!, \tag{2.6}$$

kde *k* je počet rozdílných nukleotidů u dvojice kodonů.

Př.: Kodon 1. DNA sekvence je CGA (Arg) a 2. DNA sekvence je UAC (Tyr). $P(3) = 3! = 6$, resp. při 3 změnách nukleotidů je 6 teoretických mutačních cest:

~~(1) CGA (Arg) > **UGA (Ter)** > UAA (Ter) > UAC (Tyr)~~

~~(2) CGA (Arg) > **UGA (Ter)** > UGC (Cys) > UAC (Tyr)~~
~~(3) CGA (Arg) > CAA (Gln) > **UAA (Ter)** > UAC (Tyr)~~
(4) CGA (Arg) >n> CAA (Gln) >n> CAC (His) >n> UAC (Tyr)
(5) CGA (Arg) >s> CGC (Arg) >n> UGC (Cys) >n> UAC (Tyr)
(6) CGA (Arg) >s> CGC (Arg) >n> CAC (His) >n> UAC (Tyr)

Počet cest bez terminačních kodonů je 3. Ve zbylých 3 mutačních cestách se nachází 2 synonymní a 7 nesynonymních mutací. Výpočet s_d a n_d je následující:

$$s_d = \frac{2}{3} = 0{,}667, \qquad n_d = \frac{7}{3} = 2{,}334.$$

Proporcionální vzdálenost synonymních (p_s) a nesynonymních (p_n) míst je vypočtena podílem sumy vzdáleností jednotlivých kodonů – S_d a N_d – a sumy všech možných substitucí kodonů – S a N – viz rovnice (2.7) a (2.9).

$$p_s = \frac{S_d}{S}, \tag{2.7}$$

s variancí výsledku

$$V(p_s) = \sum_{i=1}^{C}(s_{di} - p_s s_i)^2/S^2, \tag{2.8}$$

respektive

$$p_n = \frac{N_d}{N}, \tag{2.9}$$

s variancí výsledku

$$V(p_n) = \sum_{i=1}^{C}(n_{di} - p_N n_i)^2/N^2, \tag{2.10}$$

$$\mathrm{d}_s = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p_s\right), \tag{2.11}$$

s variancí výsledku

$$V(d_s) = V(p_s)/(1 - \frac{4}{3}p_s)^2, \tag{2.12}$$

respektive

$$\mathrm{d}_n = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p_n\right), \tag{2.13}$$

s variancí výsledku

$$V(d_n) = V(p_n)/(1 - \frac{4}{3}p_n)^2. \tag{2.14}$$

Pro stanovení evoluční vzdálenosti *d* se proporcionální vzdálenost přepočítává dle modelu JC nukleotidové substituce vztaženého na synonymní a nesynonymní substituce zvlášť – viz vzorce (2.11) a (2.13).

Tabulka 10: matice přechodů JC modelu, [13]

nukleotidy	**A**	**G**	**C**	**T**
A	-3α	α	α	α
G	α	-3α	α	α
C	α	α	-3α	α
T	α	α	α	-3α

Tento model je modifikovaný Markovův model evoluce, kde frekvence přechodů mezi jednotlivými nukleotidy je pro všechny stejná – viz Tabulka 10. Hodnota α vyjadřuje pravděpodobnost přechodu nukleotidů. Jedná se o jednoparametrický model. Pravděpodobnost, že jeden nukleotid přejde v jeden ze tří jiných je 3α [2], [9], [10].

2.2. Li-Wu-Luo metoda

Dvou-parametrická metoda počítající evoluční vzdálenost na základě synonymních a nesynonymních substitucích mezi dvěma DNA sekvencemi byla popsána vědci Li, Wu a Luo roku 1985 [11]. Parametry jsou založené na dvouparametrickém modelu evoluce Kimura. Tento model oproti JC modelu rozděluje změny nukleotidů na tranzici – α, které se objevují v reálných datech častěji, a transverzi – β. Model zanedbává četnost purinů a pyrimidinů v sekvenci, považuje pravděpodobnost výskytu všech nukleotidu rovnu 0,25 stejně jako u JC modelu. Obecně se matice přechodů modelu Kimura udává, jak je zobrazena v Tabulka 11. Celkový počet substitucí za čas *t* je obecně popisován jako $\alpha + 2\beta$. Byly vytvořeny vzorce na počet nukleotidů, které se mění v důsledku tranzice *P*, a počet nukleotidů, které se mění v důsledku transverze *Q*:

$$P = \frac{1}{4}(1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t}) \tag{2.15}$$

$$Q = \frac{1}{2}(1 - e^{-8\beta t}) \tag{2.16}$$

Jelikož čas *t* je neznámý, slučují se vzorce výpočtu *P* a *Q* se vzorcem (2.1) přes α a β proměnné v podobu:

$$d = 2rt = 2\alpha t + 4\beta t = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q) \tag{2.17}$$

Tabulka 11: matice přechodů Kimura modelu, [13]

nukleotidy	**A**	**G**	**C**	**T**
A	-α-2β	α	β	β
G	α	-α-2β	β	β
C	β	β	-α-2β	α
T	β	β	α	-α-2β

V rámci Li-Wu-Luo metody se rozřazují nukleotidová místa na tři třídy – ne-degenerativní (z anglického nondegenerate, někde označována jako 0-fold degenerate), dvojitě-degenerativní (z anglického twofold degenerate) a čtyřnásobně-degenerativní (z anglického fourfold degenerate). Tyto místa mají i své číselné označení $i = (0, 2, 4)$. Li, Wu a Luo [11] nevyčleňují ve své metodě mutace vedoucí k terminačním kodonům. Jejich pravděpodobnost výskytu je pouze kolem 4%, takže metoda uvažuje o terminačních kodonech jako o nesynonymní mutaci a z toho důvodu dochází k jistému zvýšení evoluční vzdálenosti d_N.

Ne-degenerativní třída je označována taková, ve které jsou všechny změny nesynonymní nebo nesmyslné. Dvojitě-degenerativní třída se definuje, jakožto cesta změn nukleotidů v kodonu, kdy dojde alespoň k jedné synonymní mutaci, ale ne všechny cesty jsou synonymní. Tedy vyjádřeno jinak – při změně nukleotidu v kodonu na některý z ostatních tří, alespoň jeden vytvoří synonymní mutaci a alespoň jeden vytvoří nesynonymní mutaci. Poslední třídou je čtyřnásobně-degenerativní, kdy dochází ke změnám jen a pouze na kodony se stejnou výslednou aminokyselinou, tedy dochází vždy k synonymním mutacím.

Příklad rozřazení do tříd pro kodon CGA pro aminokyselinu Arginin (Arg):
Čtyřnásobně-degenerativní třída – změna nukleotidu na 3. pozici v kodonu:

CGC – Arg
CGG – Arg
CGU – Arg

Dvojitě-degenerativní třída – změna nukleotidu na 1. pozici v kodonu:

AGA – Arg
GGA – Gly
UGA – terminační kodon

Ne-degenerativní třída – změna nukleotidu na 2. pozici v kodonu:

CAA – Gln
CCA – Pro
CUA – Leu

Pro samotné počítání evolučních vzdáleností jsou potřeba typy míst pro všechny kodony obou DNA zkoumaných sekvencí – označené písmenem L_i, kde i jsou jednotlivé příslušnosti do tříd. Kodony se zkoumají nukleotid po nukleotidu a zaměňují se za další tři možné nukleotidy. Pro každou pozici se určuje, o jakou třídu L_i se jedná a její počet se zvýší o jedno, obdobně, jak se počítalo S u metody Nei-Gojobori. Celkový počet L_i tříd se nakonec vydělí počtem DNA sekvencí, aby se hodnota zprůměrovala. Následně se spočítají veškeré rozdíly mezi jednotlivými kodony podle toho, zdali se jedná o tranzici – proporcionální četnost tranzicí označené písmenem P_i, či o transverzi – proporcionální četnost transverzí označené písmenem Q_i, kde i je opět příslušnost do určité třídy popsané výše. Při dvou či třech změnách nukleotidů při přechodu jednoho kodonu první sekvence DNA na druhý kodon na stejné pozici v druhé sekvenci DNA je nutné generovat pravděpodobné cesty obdobně jako u metody Nei-Gojobori. Pro stanovení pravděpodobnosti mutačních cest je možné postupovat různými způsoby. Nyní pro tuto práci se bude využívat pouze dle článku [12] faktu, že všechny změny v kodonech mají stejnou pravděpodobnost.

Když jsou připraveny veškeré proměnné L_i, P_i a Q_i, vypočtou se pomocné proměnné A_i a B_i pro finální výpočet evolučních vzdáleností d_N pro nesynonymní substituce, respektive d_S pro synonymní substituce. A_i jsou počty tranzičních substitucí založené na rozdílu proporcionálního počtu tranzicí a transverzí aproximované Kimurovým modelem – viz vzorec (2.18), po úpravě (2.22). Obdobně se počítá B_i počty transverzí – viz vzorec (2.19), po úpravě (2.24). Obě proměnné jsou opět vztaženy dle i-třídy.

$$A_i = \frac{1}{2}\ln(a_i) - \frac{1}{4}\ln(b_i) \tag{2.18}$$

$$B_i = \tfrac{1}{2}\ln(b_i), \tag{2.19}$$

kde

$$a_i = \frac{1}{1 - 2P_i - Q_i} \tag{2.20}$$

a

$$b_i = \tfrac{1}{1-2Q_i}. \tag{2.21}$$

Respektive

$$A_i = -\tfrac{1}{2}\ln(1 - 2P_i - Q_i) + \tfrac{1}{4}\ln(1 - 2Q_i), \tag{2.22}$$

s variancí výsledku

$$V(A_i) = [a_i^2 P_i + c_i^2 Q_i - (a_i P_i + c_i Q_i)^2]/L_i \tag{2.23}$$

a

$$B_i = -\frac{1}{2}\ln(1 - 2Q_i), \tag{2.24}$$

s variancí výsledku

$$V(B_i) = [b_i^2 Q_i(1 - Q_i)]/L_i. \tag{2.25}$$

Nyní již lze vypočíst evoluční distance pro synonymní (d_S) a nesynonymní (d_N) substituce – viz vzorce (2.26) a (2.28). Pro d_S se využívá do výpočtu proměnné i = 4, jelikož se jedná o substituce pouze synonymní, a částečně proměnné třídy i = 2. Dvojitě-degenerativní třída obsahuje tranziční mutace (*A2*), které jsou většinou synonymní, ale transverzální mutace jsou většinově nesynonymní. Z toho důvodu se tato třída do d_S započítává pouze z jedné třetiny. Naopak pro d_N se počítá se zbývajícími dvěma třetinami dvojitě-degenerativní třídy a celou třídou ne-degenerativní, která obsahuje pouze nesynonymní substituce.

$$d_S = 3[L_2 A_2 + L_4(A_4 + B_4)]/(L_2 + 3L_4), \tag{2.26}$$

s variancí výsledku

$$V(d_S) = 9\left[L_2^2 V(A_2) + L_4^2\left(V(A_4) + V(B_4)\right)\right] /(L_2 + 3L_4)^2 \tag{2.27}$$

a

$$d_N = 3[L_0(A_0 + B_0) + L_2 B_2]/(2L_2 + 3L_0), \tag{2.28}$$

s variancí výsledku

$$V(d_n) = 9[L_0^2(V(A_0) + V(B_0)) + L_2^2 V(B_2)] /(2L_2 + 3L_0)^2 \tag{2.29}$$

Li-Wu-Luo metoda je vhodná pro dlouhé sekvence, více jak 100 kodonů, s malou rozlišností. V takovém případě výpočty mají podobné hodnoty jako metoda předchozí, pouze přidává větší věrohodnost z důvodu přidružení více parametrů k výpočtům – rozlišování tranzičních a transverzálních mutací. Naopak při kratších sekvencích může tato metoda negativní odhad, jelikož výpočty a_i a b_i mají tendence k chybám [2], [6], [9], [11], [13].

2.3. Comeron metoda

Jak již bylo zmíněno, tato metoda rozšiřuje metodu Li-Wu-Luo. Základní změnou mezi těmito dvěma metodami je rozlišení dvojitě-degenerativní třídy, kdy alespoň jedna změna mezi dvěma triplety je synonymní, na další dvě podtřídy. Nově index *i* při této metodě nabývá označení 0, 2V, 2S a 4. Základní označení dvou tříd zůstává – nedegenerativní (i = 0) a čtyřnásobně-degenerativní (i = 4). Nové označení 2V-třídy odpovídá případům, kdy tranzice jsou nesynonymní a alespoň jedna transverze je synonymní. Naopak do 2S-třídy jsou zařazeny případy, kdy transverze jsou nesynonymní a alespoň jedna tranzice je synonymní. Tímto se zpřesňuje výpočet metody Li-Wu-Luo, kde se nadhodnocuje výpočet A_2 a podhodnocuje výpočet B_2 [2], [21].

Příklad rozřazení do 2V-třídy pro triplet CGG kódující Arginin a změny na 1. pozici:

AGG – Arginin – synonymní substituce, transverze
GGG – Glycin – nesynonymní substituce, transverze
TGG – Tryptofan – nesynonymní substituce, tranzice

Příklad rozřazení do 2S-třídy pro triplet AGT kódující Serin a změny na 3. pozici:

AGA - Arginin – nesynonymní substituce, transverze
AGC – Serin – synonymní substituce, tranzice
AGG – Arginin – nesynonymní substituce, transverze

Nadále zůstává výpočet hodnot L_i, počet míst, A_i, frekvence tranzicí. Procentuální zastoupení míst L_i je dle metody Comeron procentuálně v pořadí i = (0, 2S, 2V, 4) následovně – 64,8%, 14,5%, 2,1% a 18,6%. Pro výpočet A_i, kde i = 0 a 4, se nic nemění. Dále podléhá výpočtu dle (2.22). Pro nově vzniklé třídy je výpočet obdobný, ale vyskytují se v něm další proměnné – viz (2.30) a (2.33). Nově se počítá proměnná Q_s, frekvence synonymní transverze – viz (2.35), a Q_n, frekvence nesynonymní transverze – viz (2.37), kde q_i je proporcionální počet změn transverzí mezi sekvencemi na korespondujících místech přiřazena do *i*-třídy.

$$A_{2S} = -\frac{1}{2}\ln(1 - 2P_{2S} - Q_n) + \frac{1}{4}\ln(1 - 2Q_n) \tag{2.30}$$

s variancí

$$V(A_{2S}) = \{a_{2S}^2 P_{2S} + c_{2S}^2 V(Q_n) - [a_{2S}P_{2S} + c_{2S}V(Q_n)]\}^2 / L_{2S}, \tag{2.31}$$

kde parametr c_i je vypočten dle zadaného vzorce

$$c_i = (a_i - b_i)/2. \tag{2.32}$$

$$A_{2V} = -\frac{1}{2}\ln(1 - 2P_{2V} - Q_s) + \frac{1}{4}\ln(1 - 2Q_s) \tag{2.33}$$

s variancí

$$V(A_{2S}) = \{a_{2V}^2 P_{2V} + c_{2V}^2 V(Q_s) - [a_{2V}P_{2V} + c_{2V}V(Q_s)]\}^2/L_{2V}. \tag{2.34}$$

$$Q_s = (q_{2V} + q_4)/(L_{2V} + L_4) \tag{2.35}$$

s variancí výsledku

$$V(Q_s) = [\sum_{i=1}^{C}(q_{i2V} + q_{i4}) - E(q_{2V} + q_4)]/(L_{2V} + L_4). \tag{2.36}$$

$$Q_n = (q_0 + q_{2S})/(L_0 + L_{2S}) \tag{2.37}$$

s variancí výsledku

$$V(Q_n) = [\sum_{i=1}^{C}(q_{i0} + q_{i2S}) - E(q_0 + q_{2S})]/(L_0 + L_{2S}). \tag{2.38}$$

Pro výpočet transverzí *B* jsou zvoleny u této metody pozměněné výpočty oproti předchozí metodě – viz (2.39). Nově jsou třídy proměnné B_j pouze *j = (s, n)*, kde *s* značí synonymní transverze a *n* naopak nesynonymní transverze.

$$B_j = -\frac{1}{2}\ln(1 - Q_j) \tag{2.39}$$

s variancí výsledku B_s

$$V(B_s) = b_s^2 V(Q_s)[1 - V(Q_s)]/(L_{2V} + L_4) \tag{2.40}$$

a s variancí výsledku B_n

$$V(B_n) = b_n^2 V(Q_n)[1 - V(Q_n)]/(L_0 + L_{2S}), \tag{2.41}$$

kde proměnné b_s a b_n jsou vypočteny dle následujících vztahů

$$b_s = 1/[1 - 2V(Q_s)], \tag{2.42}$$

$$b_n = 1/[1 - 2V(Q_n)]. \tag{2.43}$$

Pro výpočet evoluční distance pro synonymní (d_s) a nesynonymní (d_n) substituce metodou Comeron se využívá součtu počtu synonymních substitucí tranzicí a transverzí respektive nesynonymních substitucí tranzicí a transverzí pro d_n – viz (2.44) respektive (2.46) [21].

$$d_s = \frac{L_{2S}A_{2S} + L_4A_4}{L_{2S} + L_4} - \frac{1}{2}\ln(1 - 2Q_s) \tag{2.44}$$

s variancí výsledku d_s

$$V(d_s) = \frac{L_{2S}^2 V(A_{2S}) + L_4^2 V(A_4)}{(L_{2S}+L_4)^2} + V(B_s). \tag{2.45}$$

$$d_n = \frac{L_{2V}A_{2V} + L_0A_0}{L_{2V} + L_0} - \frac{1}{2}\ln(1 - 2Q_n) \tag{2.46}$$

s variancí výsledku d_n

$$V(d_n) = \frac{L_{2V}^2 V(A_{2V}) + L_0^2 V(A_0)}{(L_{2V}+L_0)^2} + V(B_n). \tag{2.47}$$

2.4. Další metody

Mimo popsané metody se používají i další – např. modifikovaná metoda Nei-Gojobori, u které se zavádí 2-parametrický Kimura model pro přesnější výpočet synonymních a nesynonymních distančních vzdáleností. Popsaná klasická metoda Nei-Gojobori nadhodnocuje výpočty synonymních substitucí, jelikož zanedbává rozdíl mezi transverzí a tranzicí, což je důležité obzvlášť u třetí pozice v kodonu. Metody Pamilo-Bianchi-Li rozšiřuje metodu Li-Wu-Luo. Jelikož u Li-Wu-Luo metody bylo již zmíněno, že nadhodnocuje výpočet A_2 a podhodnocuje výpočet B_2, neboli nadhodnocuje synonymní substituce v druhé skupině a podhodnocuje nesynonymní substituce v té samé skupině. Z toho důvodu se u metody Pamilo-Bianchi-Li zavedl vážený průměr, jelikož transverze v čtyřnásobně-degenerativní třídě jsou také synonymní. Optimalizace výsledků d_s a d_n jsou následující [2]:

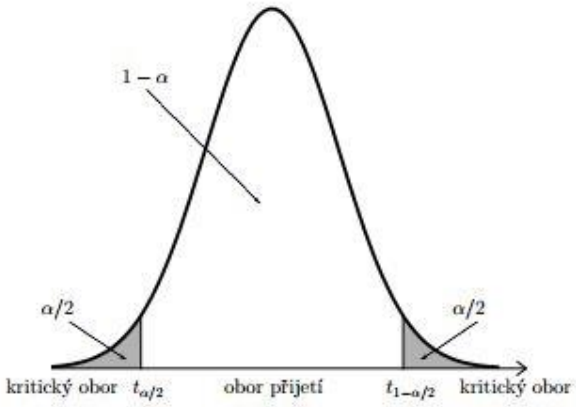
$$d_s = \frac{L_2A_2 + L_4A_4}{L_2 + L_4} - B_4 \tag{2.48}$$

$$d_n = A_0 + \frac{L_0B_0 + L_2B_2}{L_2 + L_0} \tag{2.49}$$

3. Statistické vyhodnocení výsledků

Pro zjištění pozitivní selekce je důležité statisticky dokázat, že hodnoty distancí synonymní a nesynonymní substituce jsou rozdílné, respektive v případě výpočtů selekčního tlaku dle vzorce (1.5) se musí při pozitivní selekci ($\omega > 1$) statisticky dokázat, že d_n je podstatně větší jak d_s.

Pro tento případ se využívá Z-test, jelikož známe kvalitní odhad rozptylu. Nulovou hypotézu, kterou se statistickým testem snažíme vyvrátit, se stává rovnost obou hodnot – $d_n = d_s$. Jelikož potřebujeme přijmout alternativní hypotézy, že $d_n > d_s$ u pozitivní selekce a $d_n < d_s$ u negativní selekce, purifikace, jedná se o oboustranný statistický Z-test. Z tabulky hodnot distribuční funkce *Φ(x)* normované normální náhodné veličiny – viz Příloha č. 2 – Hodnoty distribuční funkce *Φ(x)* normované normální náhodné veličiny – je zřejmé, že k zamítnutí nulové hypotézy na hladině významnosti testu $\alpha = 5\%$ je zapotřebí, aby výsledná hodnota Z-testu byla větší jak ±1,96 u oboustranného testu. Jedná se o hodnotu 0,025-kvantilu, respektive 0,975-kvantilu, jelikož α je u oboustranného testu rozdělena na obě strany – viz Obrázek 9.



Obrázek 9: Oboustranný test, [40]

Obecná formulace Z-testu – viz rovnice (3.1) – s výsledkem odpovídajícím normované normální náhodné veličině *N(0,1)*.

$$Z = (\theta - \theta_0)/s(\theta - \theta_0). \tag{3.1}$$

Pro výpočet hodnoty Z-testu při výpočtech selekčního tlaku již zmíněnými metodami byl stanoven podíl

$$Z = D/s(D), \tag{3.2}$$

kde *D* značí diferenci distančních vzdáleností

$$D = d_n - d_s \tag{3.3}$$

a *s(D)* je směrodatná odchylka rozdílu *D*, která je vypočtena jako

$$s(D) = \sqrt{V(D)}, \tag{3.4}$$

kde variance hodnot *D* – *V(D)* je součtem rozptylů

$$V(D) = V(d_n) + V(d_s). \tag{3.5}$$

Samotné výpočty variací distančních vzdáleností byly již zmíněny u každé metody zvlášť – viz vzorce (2.12) a (2.14) u metody Nei-Gojobori, (2.27) a (2.29) u metody Li-Wu-Luo a (2.45) a (2.47) u metody Comeron. Využití Z-testu koresponduje s t-testem s nekonečným stupněm volnosti [2], [48].

Mimo popisovaný Z-test se pro krátké sekvence využívá Fisherův exaktní test. Využívá se sestavení kontingenční tabulky – viz Tabulka 12, kde se zadávají hodnoty S_d. N_d, N a S. Již z této skutečnosti je zřejmé, že tento test lze využívat pouze pro Nei-Gojobori metodu. Dále je tato metoda limitovaná nízkými hodnotami a pouze reálnými čísly v tabulce, jelikož se počítají faktoriály hodnot. *T* odpovídá sumě *S* a *N*.

Tabulka 12: Fischerův exaktní test, [2]

Fisherův exaktní test			
	Synonymní místa	**Nesynonymní místa**	**Suma**
Synonymní změny	(a) S_d	(b) $S - S_d$	S
Nesynonymní změny	(c) N_d	(d) $N - N_d$	N
Suma	$S_d + N_d$	$T - S_d - N_d$	T

Pravděpodobnost získání konkrétního výsledku čtyřpolní tabulky s danými marginálními četnostmi lze vypočítat pomocí vzorce:

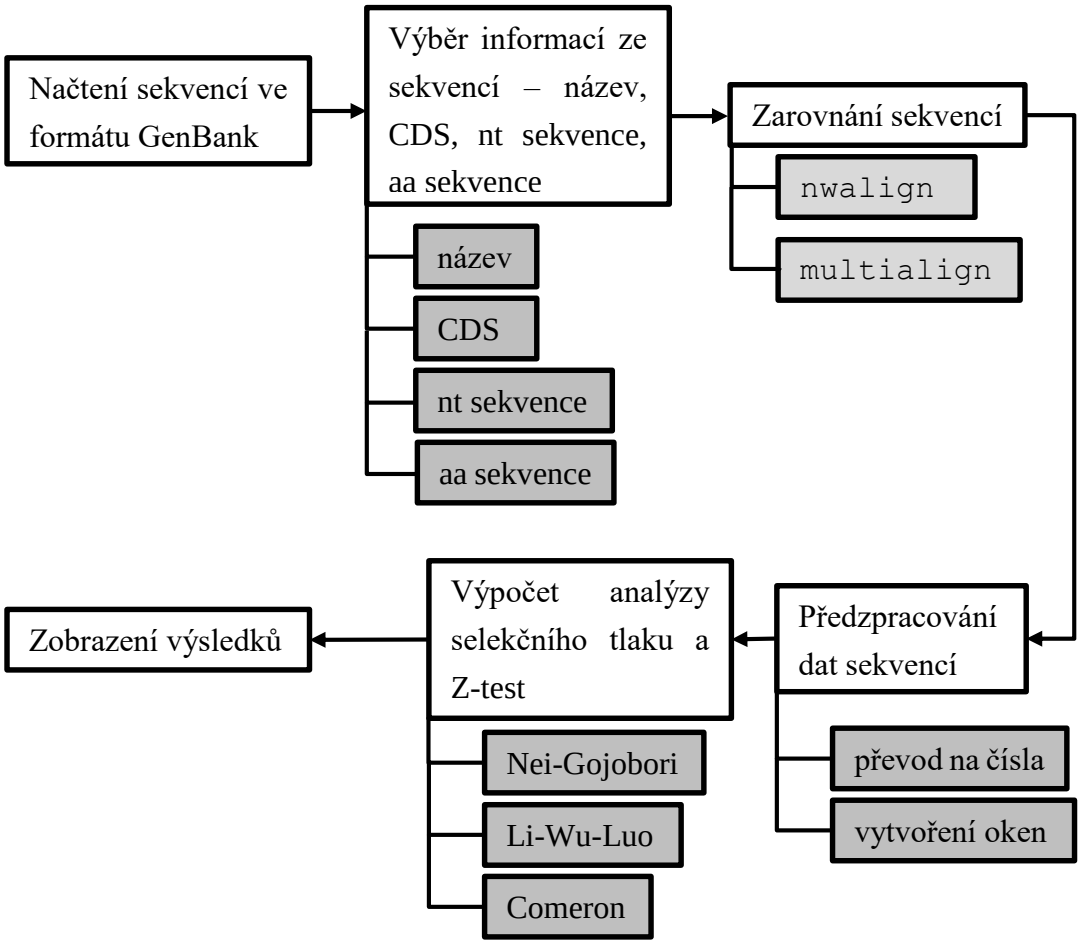
$$p = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)!(a+c)!(c+d)!(b+d)!}{n!a!b!c!d!}. \tag{3.6}$$

Hlavní myšlenkou Fisherova exaktního testu je výpočet pravděpodobnosti, se kterou bychom získali čtyřpolní tabulky stejně nebo více vzdálené od nulové hypotézy při zachování pozorovaných marginálních četností. Zachování marginálních četností znamená, že se soustředíme pouze na situace, které odpovídají stejným četnostem jednotlivých variant náhodných veličin, jako jsou v originální kontingenční tabulce [2], [44].

4. Realizace algoritmů

Pro softwarovou analýzu sekvencí byly naprogramovány funkce v programovém prostředí Matlab R2012b®. Pro uživatelsky snadnější práci s algoritmy bylo vytvořeno prostředí v Grafical User Interface (zkráceně GUI). Samotný program nese název `Slekcni_tlak.m`. Uživatelská příručka programu je připojena k práci jako příloha.

Obecné blokové schéma vytvořeného algoritmu:



4.1. Vstupní parametry k analýze

Pro zpracování analýzy programem jsou nutné sekvence z veřejných databází – např. National Center for Biotechnology Information (zkráceně NCBI). Veškeré sekvence musí být uloženy v souboru typu GenBank s koncovkou souboru *.gb . V souboru tohoto typu jsou veškeré podstatné informace, které jsou k analýze potřebné – název, nukleotidová a aminokyselinová sekvence, CDS oblasti exonů sekvence kódující určitý protein, pořadí inicializačního nukleotidu překladu.

4.2. Předzpracování sekvencí k analýze

Před samotnou analýzou, výpočtem distančních vzdáleností synonymních a nesynonymních substitucí a selekčního tlaku, je zapotřebí sekvence z GanBank souboru upravit. Z každé sekvence se používají pouze části kódující příslušný protein, takže dochází k oříznutí sekvencí dle CDS oblastí.

Takto upravené sekvence jsou následně zarovnány dle aminokyselin. Program pracuje s více druhy zarovnání. Buďto se využívá zarovnání každé sekvence s každou s využitím funkce `nwalign`, která využívá k zarovnání globální algoritmus Needleman-Wunsch [41] vzhledem parametrů – skórovací matice a sankcí za vložení mezer. Další možností zarovnání je funkcí `multialign`. Je využita skórovací matice BLOSUM (zkratka z anglického Blocks substitution matrix) [42]. Jedná se o substituční matici mezi aminokyselinami. Defaultně je nastavena na hodnotu 62 – zobrazena na Obrázek 10, která je také defaultní pro algoritmus BLAST (zkratka z anglického Basic Local Alignment Search Tool) [43]. Tato hodnota odpovídá matici vytvořené použitím sekvencí s podobností menší než 62%. Pokud se analyzují podobnější sekvence, hodnota BLOSUM matice se zadává větší. Naopak u méně podobných sekvencí se zadávají hodnoty BOLSUM matice menší. BLOSUM 62 je srovnatelná k druhé používané skórovací matici PAM160 (zkratka z anglického názvu Point accepted mutation).

```
C  9
S -1  4
T -1  1  5
P -3 -1 -1  7
A  0  1  0 -1  4
G -3  0 -2 -2  0  6
N -3  1  0 -2 -2  0  6
D -3  0 -1 -1 -2 -1  1  6
E -4  0 -1 -1 -1 -2  0  2  5
Q -3  0 -1 -1 -1 -2  0  0  2  5
H -3 -1 -2 -2 -2 -2  1 -1  0  0  8
R -3 -1 -1 -2 -1 -2  0 -2  0  1  0  5
K -3  0 -1 -1 -1 -2  0 -1  1  1 -1  2  5
M -1 -1 -1 -2 -1 -3 -2 -3 -2  0 -2 -1 -1  5
I -1 -2 -1 -3 -1 -4 -3 -3 -3 -3 -3 -3 -3  1  4
L -1 -2 -1 -3 -1 -4 -3 -4 -3 -2 -3 -2 -2  2  2  4
V -1 -2  0 -2  0 -3 -3 -3 -2 -2 -3 -3 -2  1  3  1  4
F -2 -2 -2 -4 -2 -3 -3 -3 -3 -3 -1 -3 -3  0  0  0 -1  6
Y -2 -2 -2 -3 -2 -3 -2 -3 -2 -1  2 -2 -2 -1 -1 -1 -1  3  7
W -2 -3 -2 -4 -3 -2 -4 -4 -3 -2 -2 -3 -3 -1 -3 -2 -3  1  2 11
   C  S  T  P  A  G  N  D  E  Q  H  R  K  M  I  L  V  F  Y  W
```

Obrázek 10: Skórovací matice BLOSUM62, [42]

Po zarovnání proteinových sekvencí byly dle zarovnání rozloženy i nukleotidové sekvence. Oblasti, ve kterých se při zarovnání vytvořili mezery, nebyly k výpočetní analýze využity.

Pro jednodušší programovou práci s daty jsou sekvence převedeny na typ integer, neboli IUPAC (zkratka anglického názvu organizace International Union of Pure and Applied Chemistry) písmenný zápis je převeden na čísla dle Tabulka 13.

Tabulka 13: IUPAC kódové označení a jejich číselný ekvivalent

Nukleotid	IUPAC kódové označení	Číselné označení
Adenin	A	1
Cytosin	C	2
Guanin	G	3
Thymin	T	4
Uridin (u RNA)	U	4
Puriny (A a G)	R	5
Pyrimidiny (T a C)	Y	6
Keto-skupina (G a T)	K	7
Amino-skupina (A a C)	M	8
Silná vazba (G a C)	S	9
Slabá vazba (A a T)	W	10
Není A	B	11
Není C	D	12
Není G	H	13
Není T / U	V	14
Jakýkoliv nukleotid	N	15
Mezera	-	16
Neznámý	*	0

4.3. Výpočet selekčního tlaku

Pro samotnou analýzu jsou již data připravená – načtená, zarovnaná a převedena na číselné vyjádření nukleotidů. V této kapitole se popíše princip algoritmů na výpočet selekčního tlaku, respektive parametrů d_n a d_s pro výpočet selekce.

Sekvence se prochází buď jako celek nebo v okně o velikosti, která je předem nastavena dle zadání, s krokem 1 aminokyseliny, respektive 3 nukleotidů. Každá vybraná metoda, Nei-Gojobori, Li-Wu-Luo či Comeron, má na vstupu dvojici analyzovaných sekvencí nebo jejich část vyříznutou dle nastaveného okna. Dále pro menší náročnost je na vstupu do metod i stejná část nukleotidového zápisu převedená na čísla. Mimo tyto parametry do metod vstupuje i typ genetického kódu a implementovaná databáze proměnných.

Díky implementované databázi proměnných se při každém výskytu nepočítají stejné kodony opakovaně. Pomocným skriptem je docíleno toho, že všechny možné triplety ve všech

genových kódech mají předem propočtené parametry. Jelikož máme 64 možných tripletů – 4 možné nukleotidy na první pozici, 4 možné nukleotidy na druhé pozici a 4 možné nukleotidy na třetí pozici v tripletu – a 17 různých genetických kódů, byly vytvořeny struktury, ve kterých se dohledává potřebný výsledek indexací. Pro metodu Nei-Gojobori je předpřipraveno 1088 hodnot proměnné *s*. Pro metody Li-Wu-Luo a Comeron je vypočteno 3264 hodnot parametru L_i se třemi stupni, respektive 4352 hodnot pro parametr L_i se čtyřmi stupni. Výpočet hodnot L_i byl zjednodušen faktem, že k metodě Comeron bylo zapotřebí pouze rozlišit dvojitě-degenerativní třídu vypočtenou již k metodě Li-Wu-Lou na další dvě podtřídy, jinak hodnoty zůstávaly stejné.

Pro indexaci u předem vypočtených hodnot (*s* a L_i) se využívají pozice v tripletu nukleotidů a jejich ekvivalentní číselný zápis. Tudíž vyhledávání *s* hodnot je následující a L_i obdobné:

```
datab_hodnot_s{první pozice v triplet, genetický kód}(druhá
        pozice v tripletu, tretí pozice v tripletu)
```

Dále se v databázi nachází prostor pro ukládání vypočtených hodnot pro proměnné S_d, N_d u Nei-Gojobori metody a pro uložení hodnot *P* a *Q* vypočítaných u metod Li-Wu-Luo a Comeron. Jelikož tyto hodnoty jsou podmíněné již dvojicí tripletů mezi sekvencemi, je tento prostor před prvním využitím prázdný. Je pouze připravena struktura na ukládání v průběhu analýz. Jelikož se mnohé dvojice tripletů v průběhu využití programu vyskytují několikrát, tímto krokem se docílí toho, že se již nepočítají opakovaně, pouze se hodnoty vyhledají v databázi výsledků. Všech možných kombinací je 69632 (64 tripletů jedné sekvence ku 64 tripletů druhé sekvence a 17 typů genetických kódů), z toho důvodu již nebyly tyto hodnoty vypočteny předem. Databáze se plní hodnotami postupně, dle vstupních analýz od uživatelů. Pro jednodušší indexaci v tomto případě byly jednotlivé možné triplety indexovány a v databázi, která má podobu 3D matice, se již vyhledává pouze pomocí indexů dvojic tripletů a genetického kódu.

Následně se již dle zadaných vzorců dle metod vypočtou hodnoty d_s a d_n, jejich podíl v podobě výsledku selekčního tlaku ω a dle variancí všech proměnných z-hodnota statistického testu.

4.4. Zobrazení výsledků

Pro zobrazení výsledků bylo využito tabulek nebo grafů. Záleží na typu zvolené analýzy. Při využití srovnání distancí synonymních a nesynonymních míst či selekčního tlaku více sekvencí naráz bez využití průchodu okna, se výsledky zobrazují v tabulkách srovnávající sekvence vzájemně. Pokud je využito analýzy se zadáním velikosti okna, jsou výsledky selekčního tlaku a distancí synonymních a nesynonymních míst zobrazeny v grafu, kde x-ová souřadnice odpovídá pozici okna v sekvenci po zarovnání a y souřadnice udává velikost

selekčního tlaku popřípadě distancí. Každá metoda výpočtu je zobrazena jinou barvou v grafu. Dle Z-testu jsou zeleně zvýrazněná místa, kde výsledek je signifikantní a dochází se statistickou pravděpodobností na hladině testu 5% k pozitivní či negativní selekci. Dále je vyhodnoceno procentuální vyhodnocení jednotlivých výsledků selekce dle jednotlivých metod.

5. Data

K analýze selekčního tlaku bylo využito dvou datových setů DNA stažených z veřejné databáze NCBI ve formátu GenBank. V rámci práce se analyzovaly DNA různých genových kódů. Využil se standardní genetický kód a mitochondriální kód obratlovců.

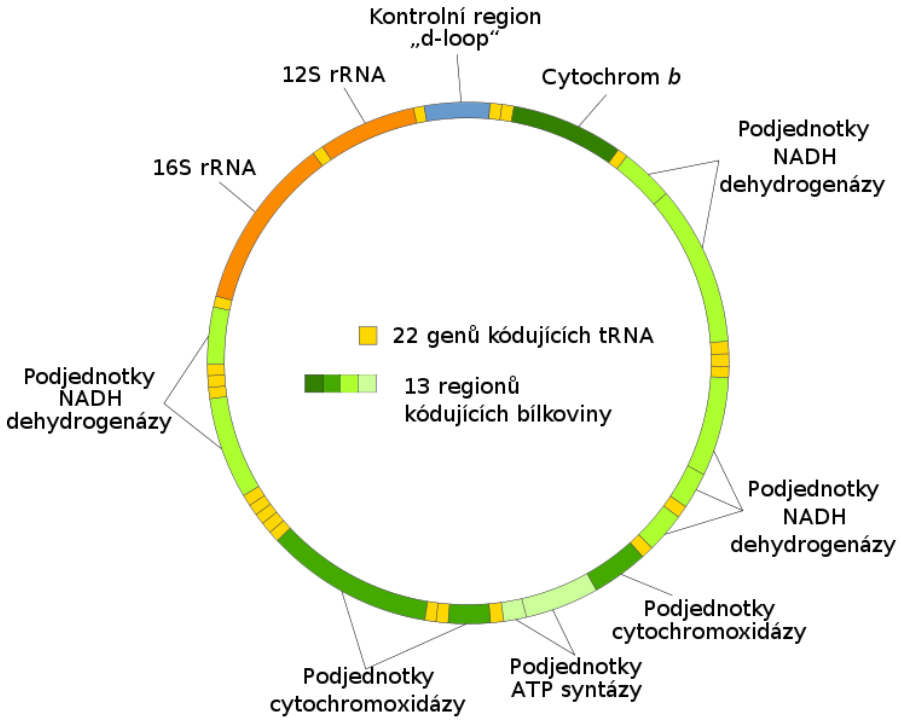
Prvním analyzovaným datovým setem jsou sekvence DNA kódující gen BRCA1. V roce 1990 byla náchylnost genu pro rakovinu prsu mapována pomocí genetické vazby na dlouhém rameni chromozomu 17, v intervalu 17q12-21. Spojení mezi rakovinou prsu a genetických markrů na chromozomu 17q byla brzy potvrzena a byl pozorován důkaz pro přenos náchylnosti rakoviny prsu a vaječníků. BRCA1 gen byl následně identifikován a bylo zjištěno, že obsahuje 24 exonů, které kódují protein o velikosti 1863 aminokyselin. Zárodečné mutace v BRCA1 jsou spojeny s rakovinou prsu, rakoviny vaječníků a rakoviny vejcovodu. Rakoviny mužského prsu, rakovina pankreatu, rakovina varlat a rakoviny prostaty může být také spojena s mutací BRCA1. Nicméně mužská rakovina prsu, rakovina pankreatu, rakovina prostaty jsou silněji spojeny s mutacemi v BRCA2 [45].

Tabulka 14: Sekvence vybrané do datového setu genu BRCA1

Druh latinsky	**Druh česky**	**Přístupové číslo GenBank**	**Délka sekvence**
Homo sapiens	Člověk moudrý	NM_007294.3	7224 bp
Bos taurus	Tur domácí	NM_178573.1	5550 bp
Canis lupus familiaris	Pes domácí	U50709.1	5637 bp
Mus musculus	Myš domácí	U35641.1	5538 bp
Rattus norvegicus	Potkan obecný	AF036760.1	5607 bp

Pro mitochondriální kód byl vytvořen datový set celých mitochondriálních DNA (zkráceně mtDNA) vybraných jedinců – savců. Jedná se o specifický druh DNA jednoduchá na izolaci, relativně malá (lidská mtDNA má velikost 16,6 kb) a cirkulárního typu. Struktura a genové uspořádání mtDNA je mezi savci zachovávána – viz Obrázek 11. Kóduje 2 rRNA geny pro mitochondriální ribozomy (12S rRNA a 16S rRNA), 22 typů tRNA a mRNA pro syntézu 13 proteinů − 3 podjednotek cytochromoxidázy, 2 podjednotky ATP syntázy, 7 podjednotek

NADH dehydorgenázy a cytochrom b. Všechny proteiny kódované mtDNA jsou součástí enzymového komplexu oxidativní fosforylace [47].



Obrázek 11: Savčí cirkulární mtDNA s vyznačením umístění některých genů, [46]

Pro analýzu mtDNA byla zvolena část u všech sekvencí, která kóduje cytochrom b. Jedná se o centrální katalytickou podjednotku chinolinu. Podílí se na vazbě chinonového substrátu a je zodpovědný za přenos trans-membránových elektronů. Cytochrom b také obsahuje oblasti, na které se váží různé inhibitory a chinonové antagonisté a inhibují oxidoreduktázu [47].

Tabulka 15: Sekvence vybrané do datového setu mtDNA

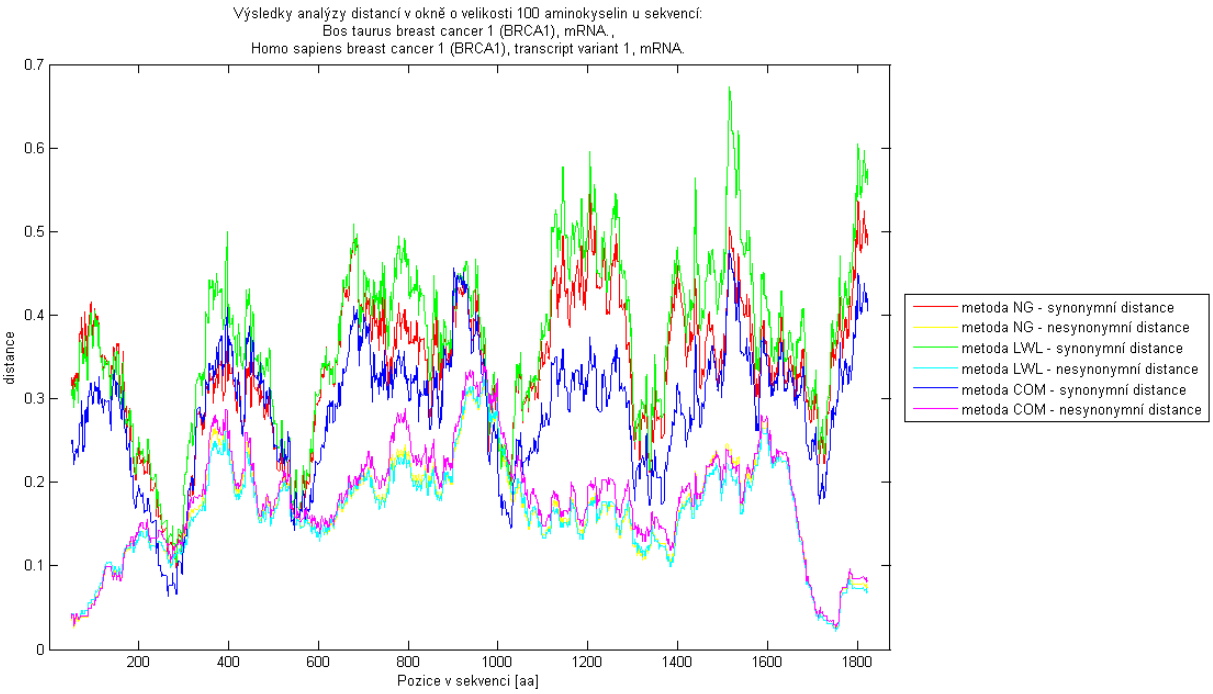
Druh latinsky	**Druh česky**	**Přístupové číslo GenBank**	**Délka sekvence**
Homo sapiens	Člověk moudrý	NC_012920.1	16569 bp
Homo sapiens neanderthalensis	Neandertálec	NC_011137.1	16565 bp
Pan paniscus	Šimpanz bonobo	NC_001644.1	16563 bp
Pan troglodytes	Šimpanz učenlivý	NC_001643.1	16554 bp
Mus musculus	Myš domácí	JQ003190.1	16300 bp
Rattus norvegicus	Potkan obecný	AC_000022.2	16300 bp
Canis lupus familiaris	Pes domácí	KF907307.1	16727 bp

6. Diskuze výsledků

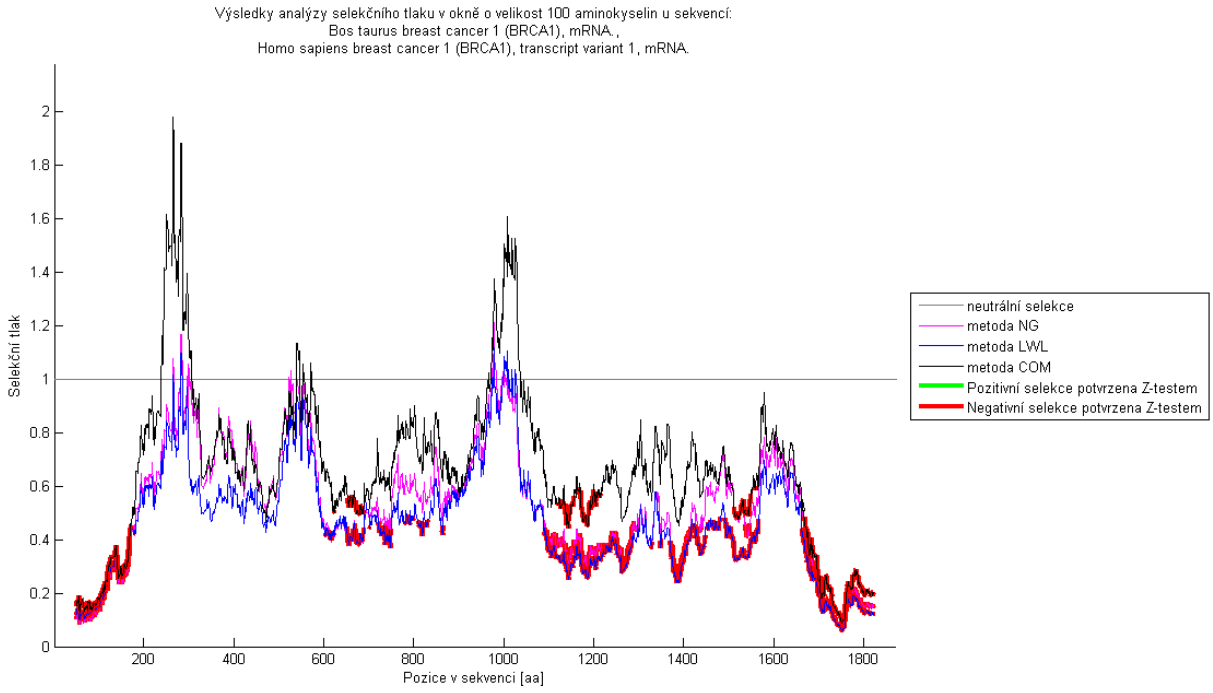
K analýze v prvním případě se využil již popsaný datový set genu BRCA1. Zobrazeny jsou grafy výsledků analýz selekčního tlaku sekvence Homo sapiens a ostatních sekvencí, na kterých se budou demonstrovat výsledky. Zbytek sekvencí bude zhodnoceno pouze statisticky.

Analýza sekvencí BRCA1 byla provedena se vstupními parametry velikostí standardního okna 100 aminokyselin, 50 aminokyselin a 200 aminokyselin, `nwalign` zarovnání se skórovací maticí BLOSUM62 a analýze podléhaly celé kódující sekvence. Byl využit standardní genetický kód. Výsledné grafy selekčního tlaku a distancí synonymních a nesynonymních substitucí budou zobrazeny u sekvence organismu *Homo sapiens* a *Bos taurus*.

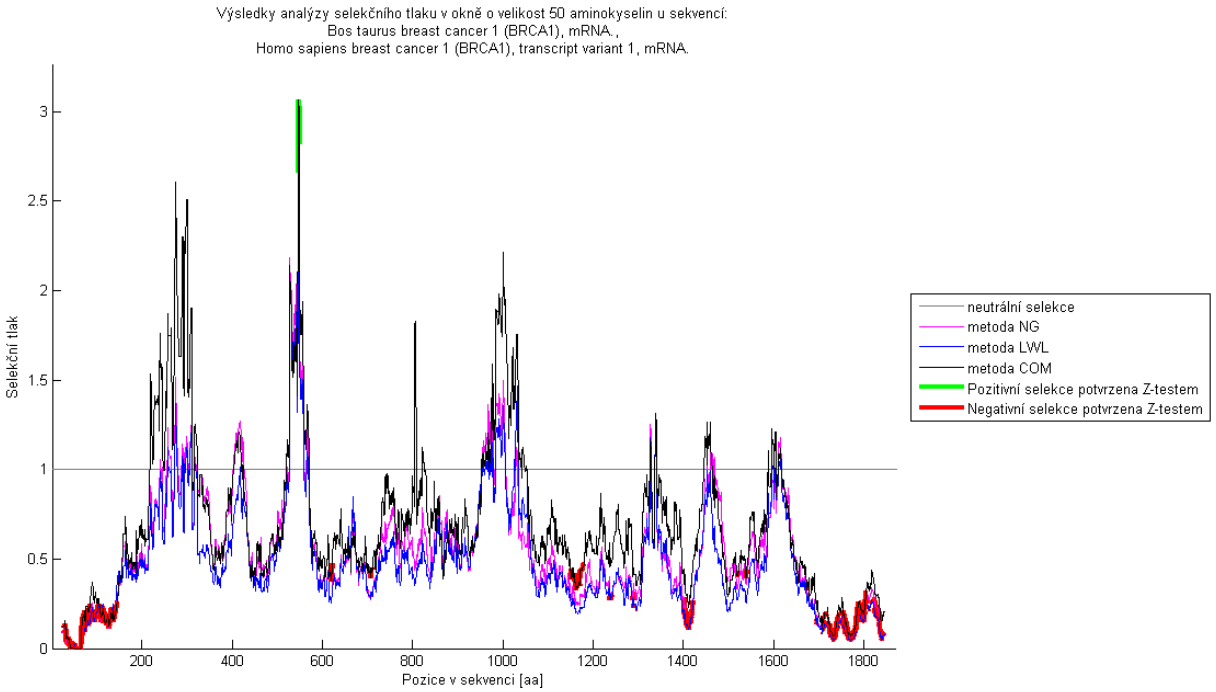
Grafické zobrazení synonymních a nesynonymních distancí jsou zobrazeny na Obrázek 12. Výsledky nesynonymních distancí kopírují vzájemně svůj vývoj. U metody Comeron je vyšší než u ostatních dvou metod. Naopak u výsledků synonymních distancí metody Comeron nabývá nižších hodnot. Li-Wu-Luo metoda vyhodnocuje synonymní distanci vyšší než ostatní metody. Tento trend byl pozorován u všech výsledků. Je to podložené i teoriemi metod. Li-Wu-Luo metody oproti Nei-Gojobori bere v úvahu terminační kodóny ve výpočtech možných evolučních cest jako nesynonymní, což zvyšuje hodnotu nesynonymní distanci. U metody Comeron je pozorována nesynonymní distance nejnižší, což je způsobeno lepším rozlišením dvojitě-degenerativní skupiny na dvě podskupiny. Na Obrázek 13, Obrázek 14 a Obrázek 15 jsou zobrazeny hodnoty podílů distancí, respektive výsledný selekční tlak s vyznačenými místy odpovídající pozitivní a negativní selekci. Pro porovnání jsou zobrazeny výsledky stejné dvojice sekvencí při různé velikosti okna – standardních 100 aa, 50 aa a 200 aa. S větší velikostí okna se výsledné hodnoty selekčního tlaku snižují. Statistickým testem dochází k větší detekci negativní selekce, jelikož se průměruje dlouhá část sekvence. Naopak s malým oknem 50 aa dochází k výsledkům, které detekují vyšší hodnoty selekčního tlaku, směřují k pozitivní selekci, ale statisticky signifikantní výsledky jsou pouze v oblasti mezi 500 a 600 aa, kde se pouze u takto malého okna objevila část podléhající pozitivní selekci. Všechny tři metody mají vývoj grafů výsledků selekčního tlaku podobný. Metoda Comeron vykazuje vyšší hodnoty selekčního tlaku oproti zbylým metodám. Metoda Li-Wu-Luo naopak vykazuje nižší hodnoty selekčního tlaku oproti ostatním metodám. Tento trend je potvrzen i na grafech výsledků selekčního tlaku pro dvojici sekvencí *Homo sapiens* a *Mus musculus* - Obrázek 16 s velikostí okna 50 aa, Obrázek 17 s velikostí okna 100 aa a Obrázek 18 s velikostí okna 200 aa.



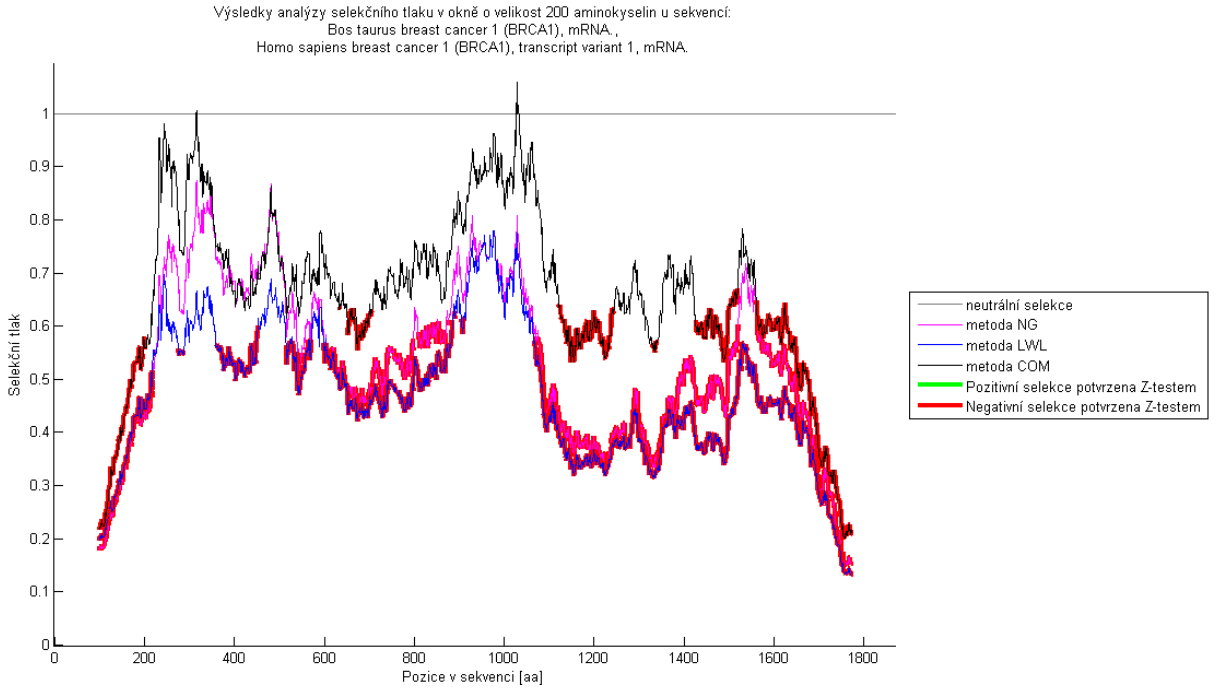
Obrázek 12: Výsledky synonymních a nesynonymních distancí mezi *Homo sapiens* a *Bos taurus* – gen BRCA1, velikost okna 100 aa



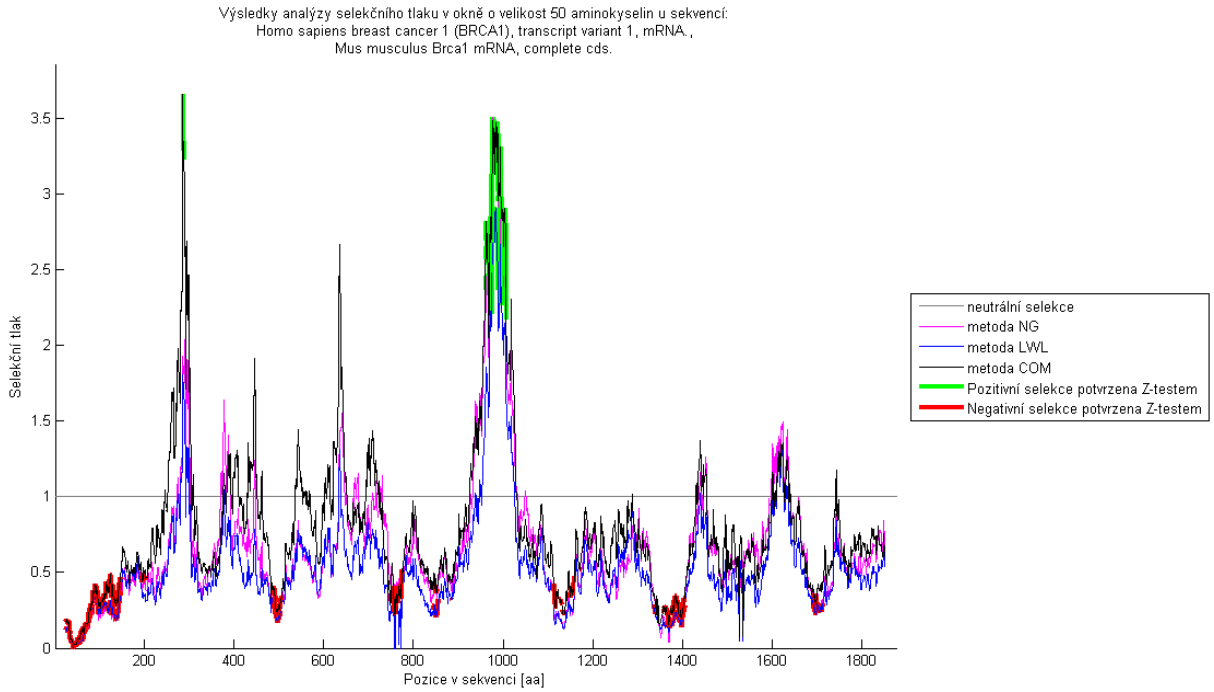
Obrázek 13: Výsledky selekčního tlaku mezi *Homo sapiens* a *Bos taurus* – gen BRCA1, velikost okna 100 aa



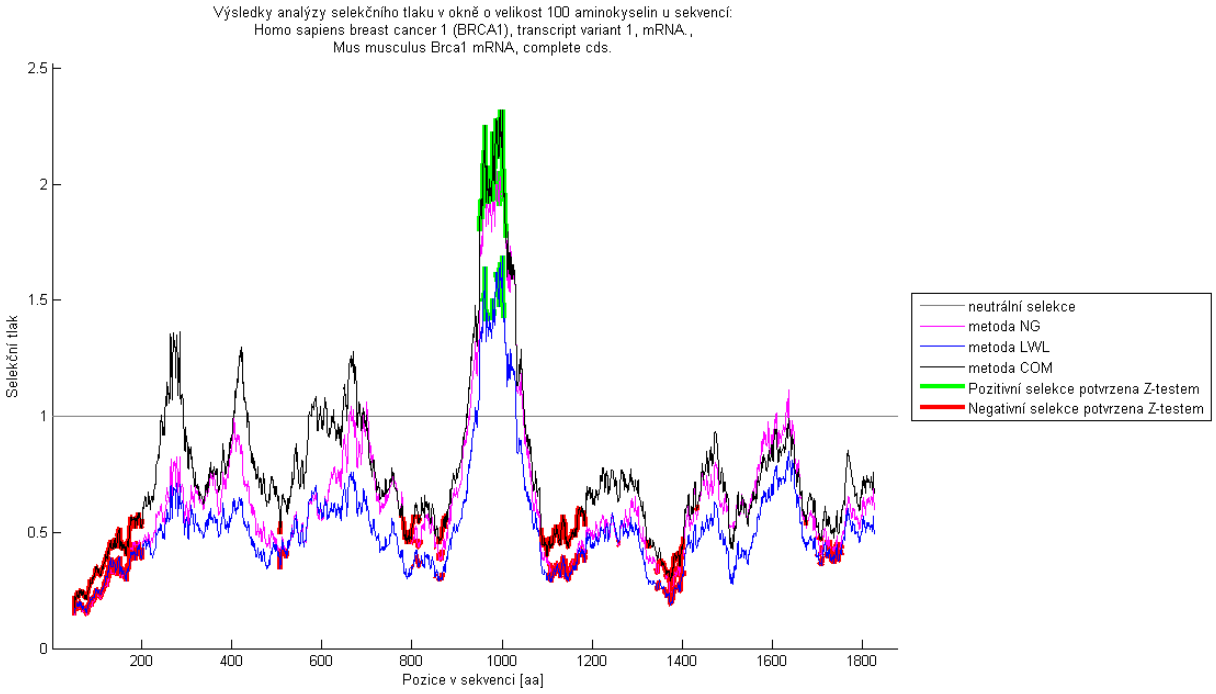
Obrázek 14: Výsledky selekčního tlaku mezi *Homo sapiens* a *Bos taurus* – gen BRCA1, velikost okna 50 aa



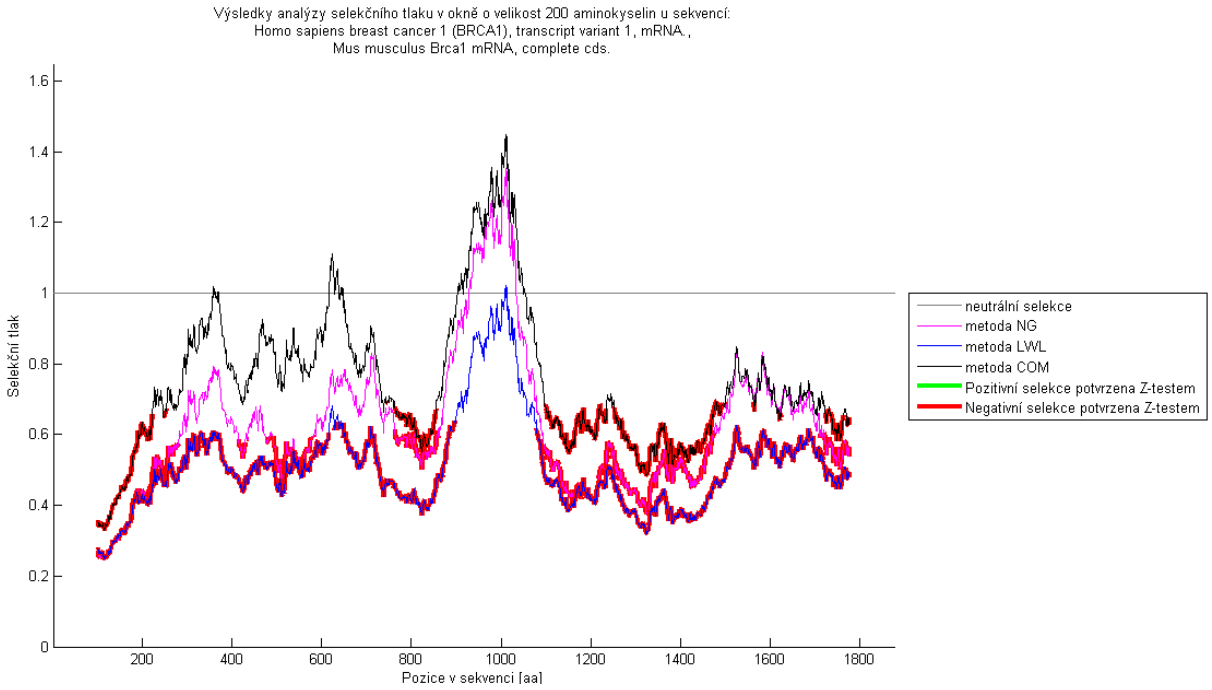
Obrázek 15: Výsledky selekčního tlaku mezi *Homo sapiens* a *Bos taurus* – gen BRCA1, velikost okna 200 aa



Obrázek 16: Výsledky selekčního tlaku mezi *Homo sapiens* a *Mus musculus* – gen BRCA1, velikost okna 50 aa

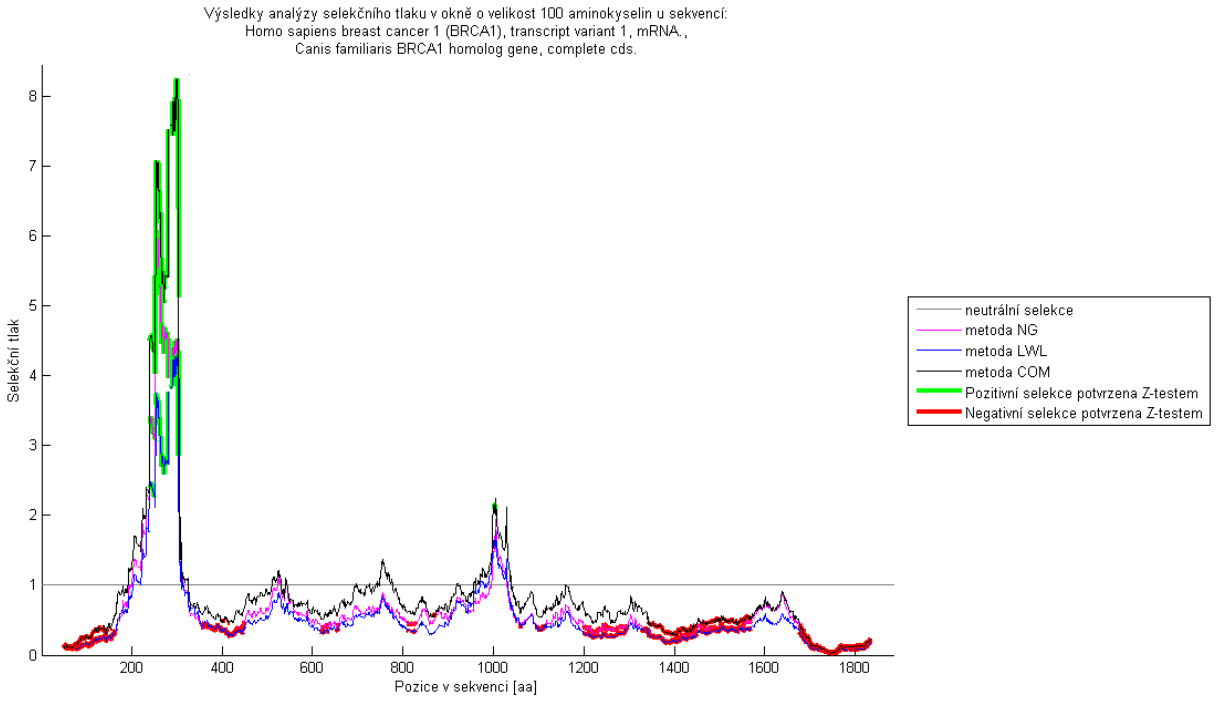


Obrázek 17: Výsledky selekčního tlaku mezi *Homo sapiens* a *Mus musculus* – gen BRCA1, velikost okna 100 aa

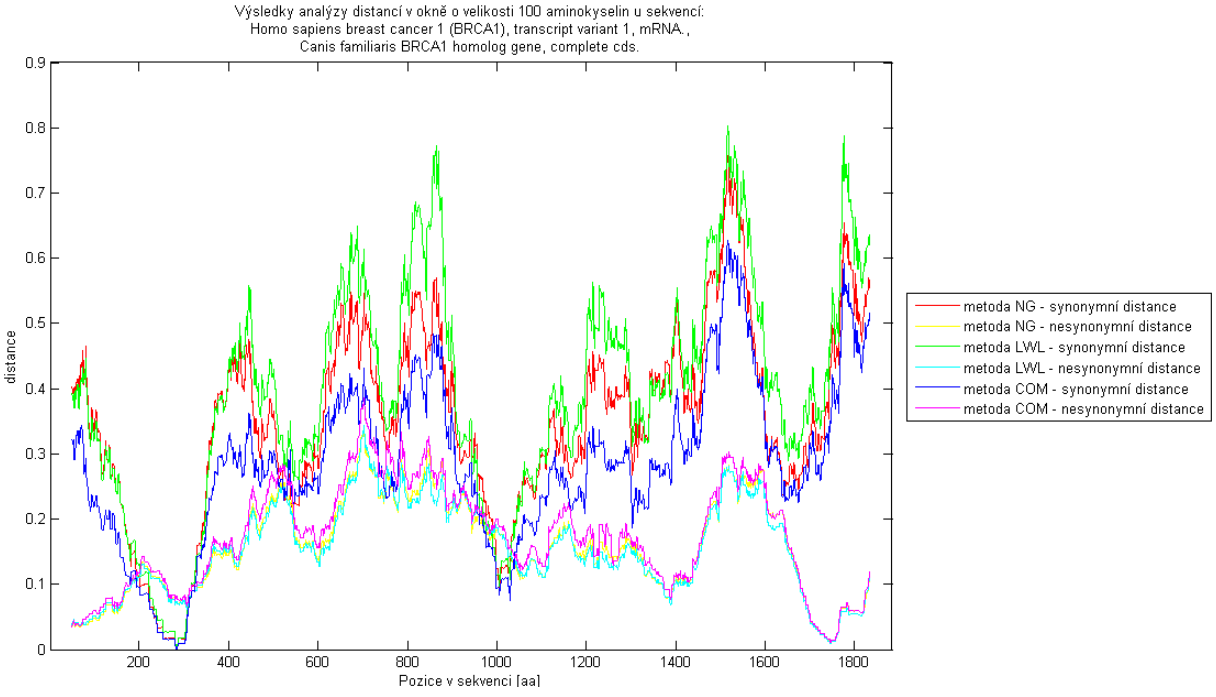


Obrázek 18: Výsledky selekčního tlaku mezi *Homo sapiens* a *Mus musculus* – gen BRCA1, velikost okna 200 aa

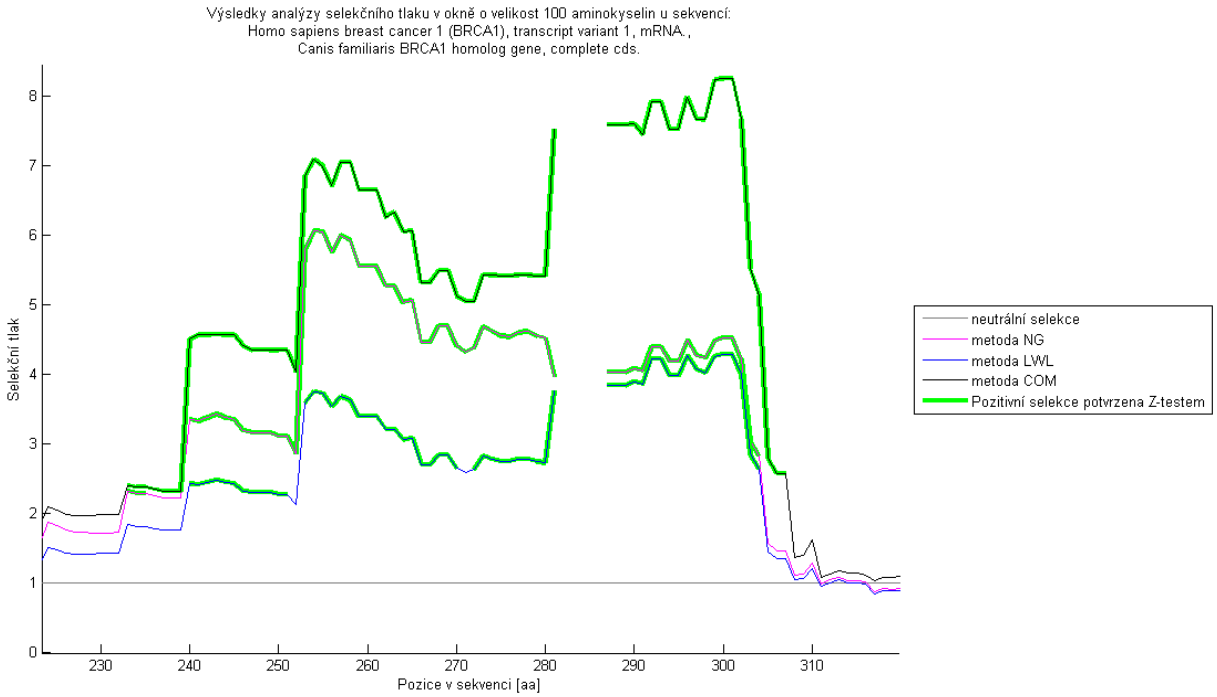
Při výpočtech selekčního tlaku byly nalezeny místa, kde se nemohl vypočítat, jelikož ze vztahu pro selekční tlak – viz rovnice (1.4) – je zřejmé, že nelze dělit 0. Pokud metody vypočítaly nulovou hodnotu pro synonymní distanci d_S, nebyl v takovém případě žádný výsledek selekce pro takovou oblast. Tato skutečnost lze vidět u výsledků selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* s parametrem velikosti okna 100 aa v oblasti se středem okna mezi 230 a 310 aa – viz Obrázek 19 a výřez oblasti na Obrázek 21. Na Obrázek 20, respektive výřezu na Obrázek 22, jsou zobrazeny výsledné hodnoty distancí, kde výsledky synonymních distancí všech metod vychází 0. Při změně velikosti okna dochází k tomu, že při zmenšení okna je oblast bez výsledků selekčního tlaku větší – viz Obrázek 23 a výřez oblasti Obrázek 24, naopak při větším okně tento jev nebyl zaznamenán – viz Obrázek 25 a výřez oblasti Obrázek 26. Tímto bylo dokázáno, že pokud se jedná o problém, kde synonymní substituce není vypočtena, ani jedna metoda nevyhodnotila situaci odlišně.



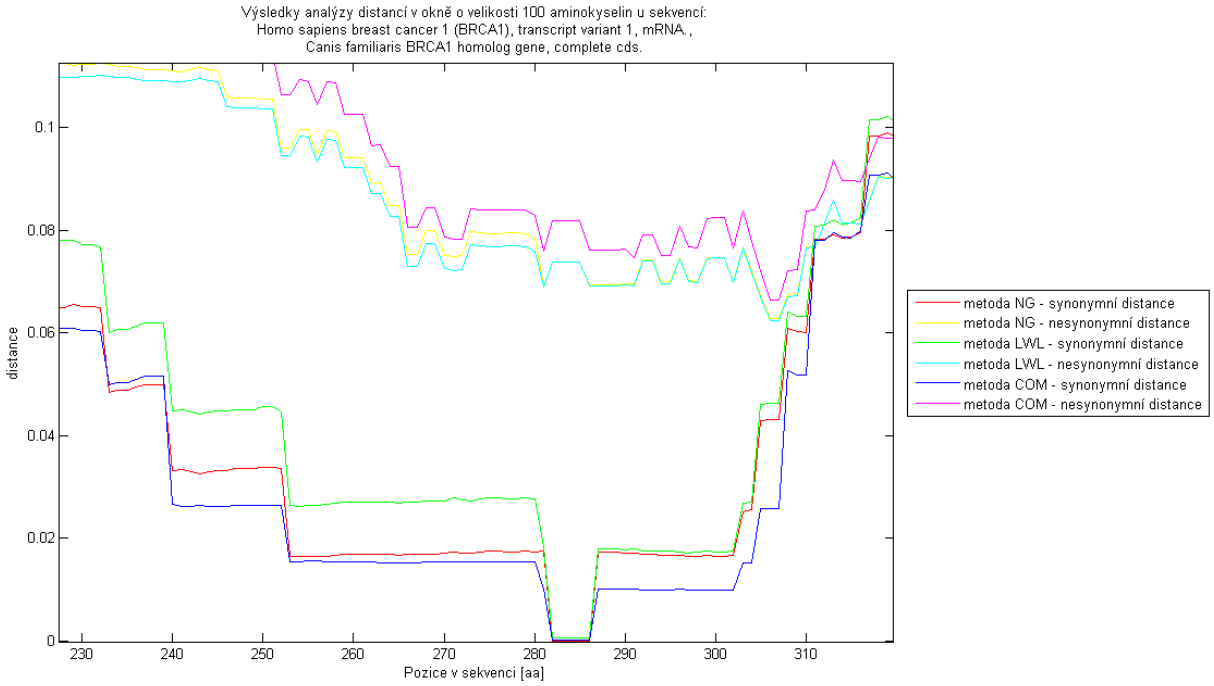
Obrázek 19: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 100 aa



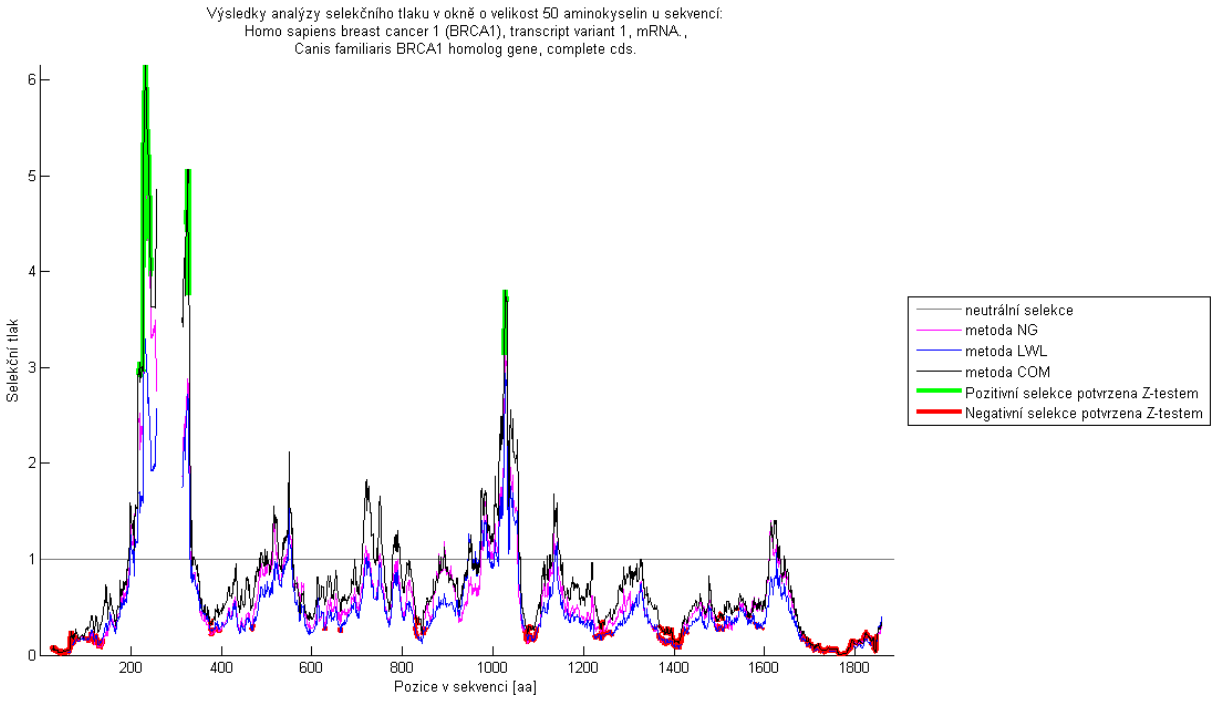
Obrázek 20: Výsledky synonymních a nesynonymních distancí mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 100 aa



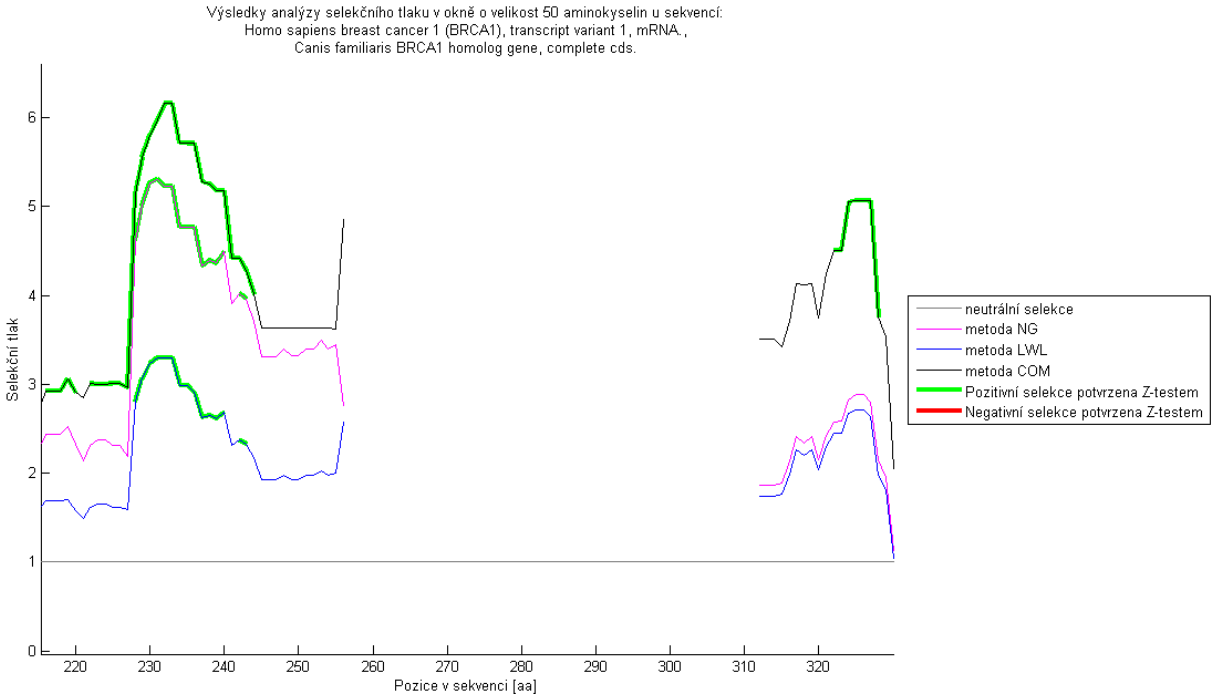
Obrázek 21: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 100 aa, výřez oblasti 230 – 320 aa



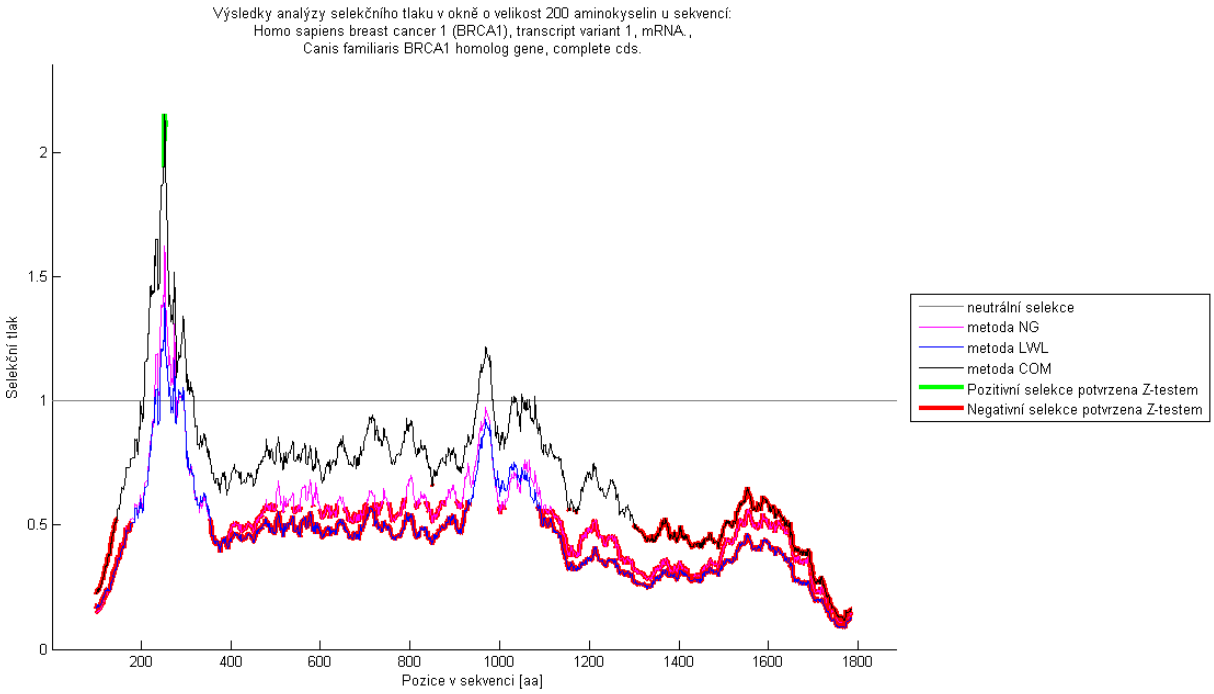
Obrázek 22: Výsledky synonymních a nesynonymních distancí mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 100 aa, výřez oblasti 230 – 320 aa



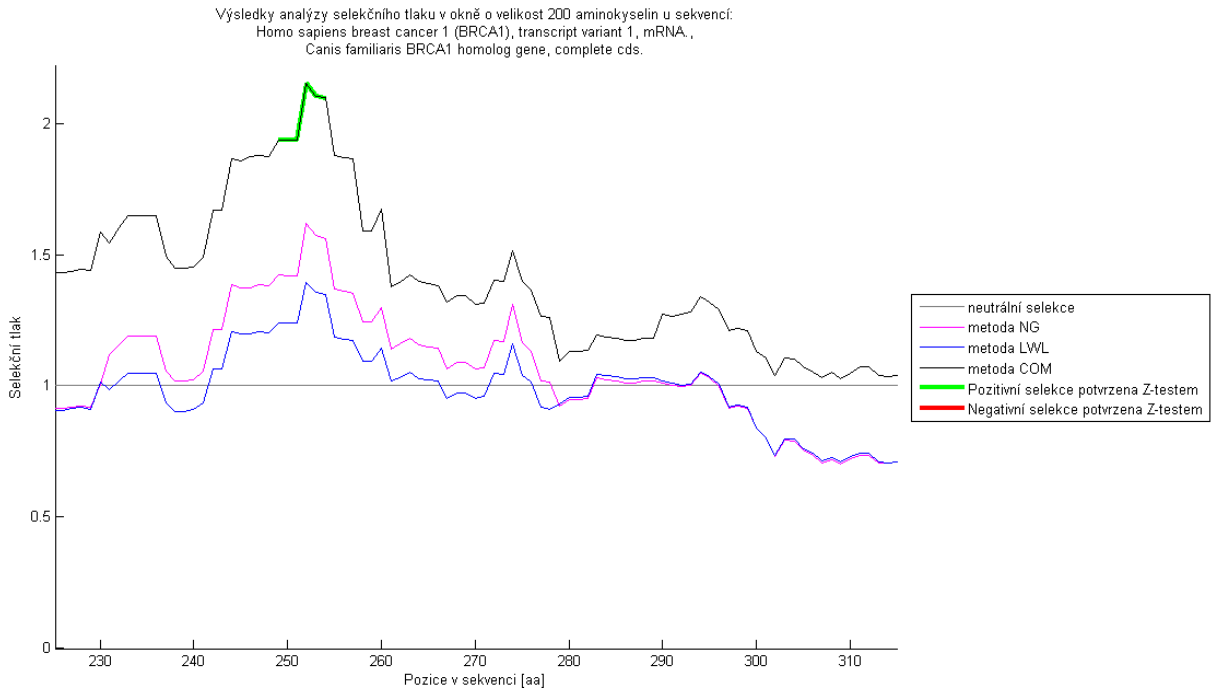
Obrázek 23: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 50 aa



Obrázek 24: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 50 aa, výřez oblasti 230 – 320 aa



Obrázek 25: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 200 aa



Obrázek 26: Výsledky selekčního tlaku mezi *Homo sapiens* a *Canis familiaris* – gen BRCA1, velikost okna 200 aa, výřez oblasti 230 – 320 aa

Pro vyhodnocení věrohodnosti výsledků jednotlivých metod s různými velikostmi okna (50, 100 a 200 aa) byla procentuální zastoupení oblastí, kde docházelo k pozitivní, respektive negativní selekci, zanesena do tabulky a zobrazena graficky – viz následující přílohy.

Tabulka 16: Procentuální zastoupení oken s detekcí pozitivní selekce u datového setu BRCA1 [%]

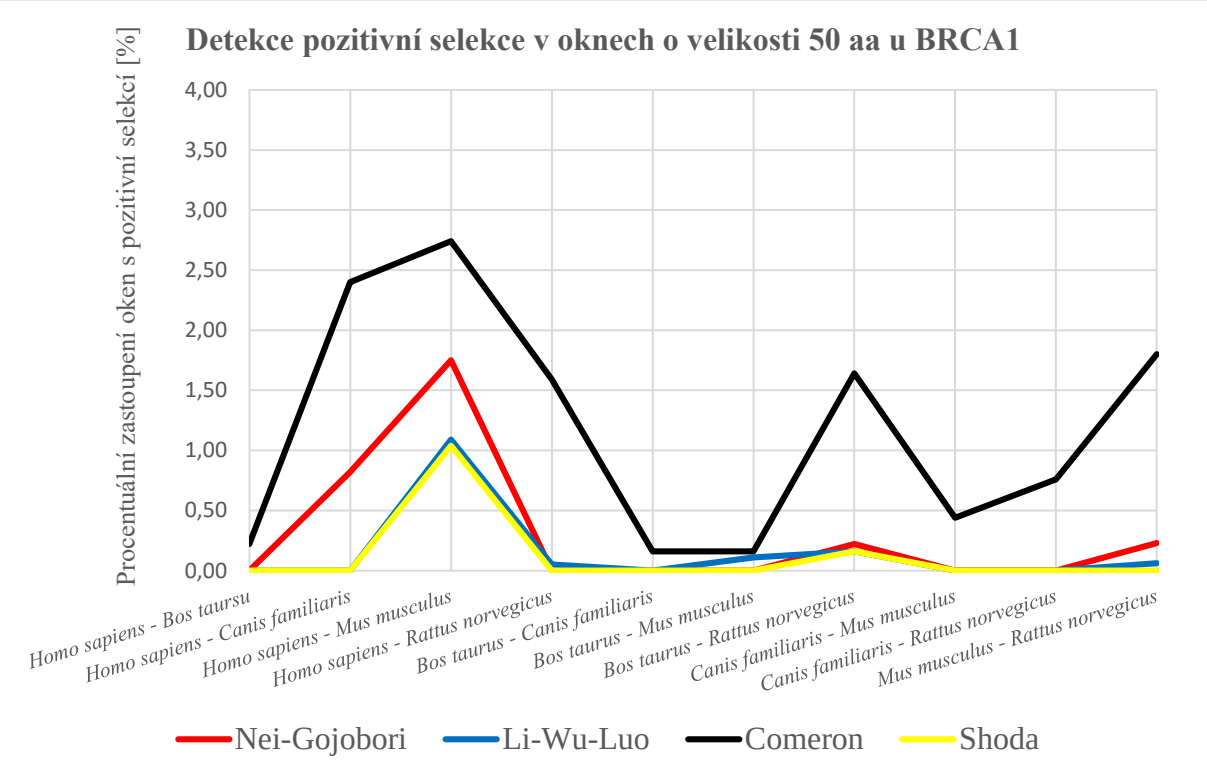
	Homo sapiens - Bos taursu	*Homo sapiens - Canis familiaris*	*Homo sapiens - Mus musculus*	*Homo sapiens - Rattus norvegicus*	*Bos taurus - Canis familiaris*	*Bos taurus - Mus musculus*	*Bos taurus - Rattus norvegicus*	*Canis familiaris - Mus musculus*	*Canis familiaris - Rattus norvegicus*	*Mus musculus - Rattus norvegicus*
	Velikost okna 50 aa									
Nei-Gojobori	0,00	0,82	1,75	0,00	0,00	0,00	0,22	0,00	0,00	0,23
Li-Wu-Luo	0,00	0,00	1,09	0,05	0,00	0,11	0,16	0,00	0,00	0,06
Comeron	0,22	2,40	2,74	1,59	0,16	0,16	1,64	0,44	0,76	1,80
Shoda	0,00	0,00	1,04	0,00	0,00	0,00	0,16	0,00	0,00	0,00
	Velikost okna 100 aa									
Nei-Gojobori	0,00	3,25	1,86	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Li-Wu-Luo	0,00	1,51	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Comeron	0,00	3,75	3,32	0,90	0,00	0,00	1,12	0,00	0,56	0,00
Shoda	0,00	1,51	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Velikost okna 200 aa									
Nei-Gojobori	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Li-Wu-Luo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Comeron	0,00	0,36	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Shoda	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Tabulka 17: Procentuální zastoupení oken s detekcí negativní selekce u datového setu BRCA1 [%]

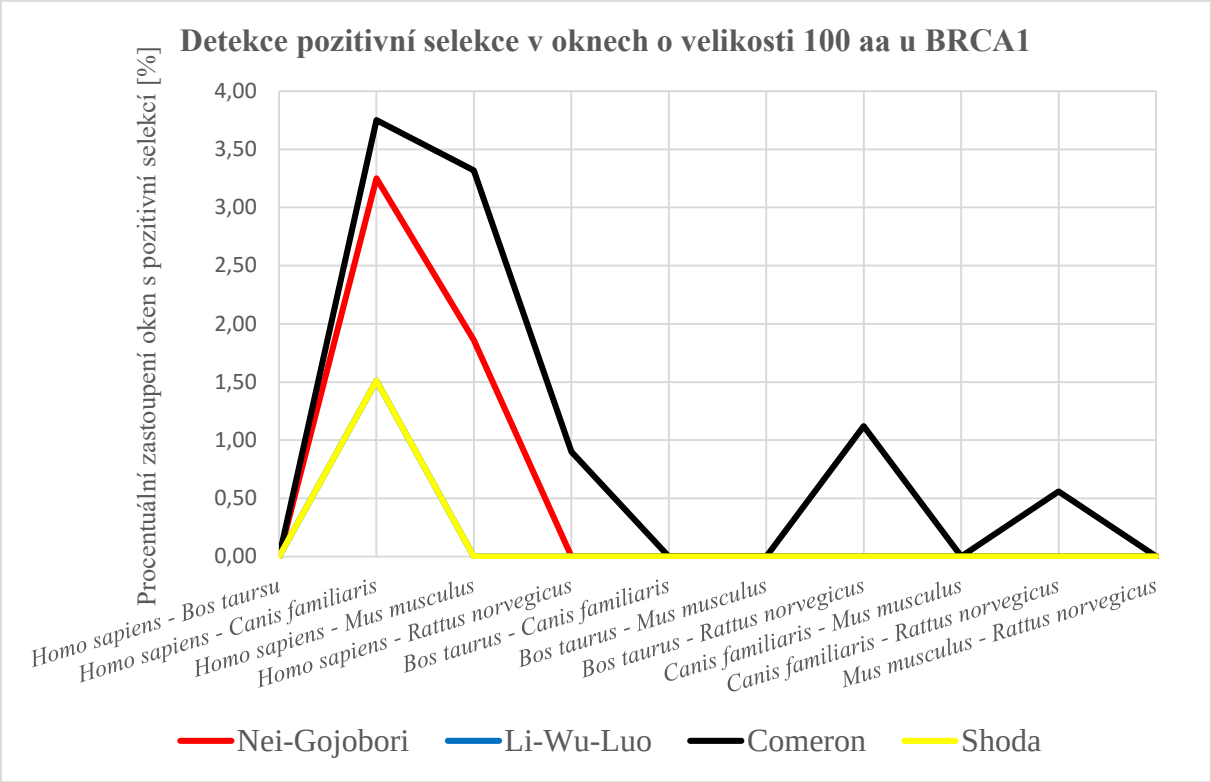
	Homo sapiens - Bos taursu	*Homo sapiens - Canis familiaris*	*Homo sapiens - Mus musculus*	*Homo sapiens - Rattus norvegicus*	*Bos taurus - Canis familiaris*	*Bos taurus - Mus musculus*	*Bos taurus - Rattus norvegicus*	*Canis familiaris - Mus musculus*	*Canis familiaris - Rattus norvegicus*	*Mus musculus - Rattus norvegicus*
	Velikost okna 50 aa									
Nei-Gojobori	13,27	20,44	4,70	6,67	18,96	4,65	6,38	7,24	9,36	20,74
Li-Wu-Luo	17,11	26,38	12,09	9,62	21,85	13,94	11,03	13,72	16,59	22,32
Comeron	12,95	15,53	15,15	9,13	17,71	17,16	15,34	20,42	22,36	18,21
Shoda	7,63	12,86	4,27	3,45	11,66	3,50	3,93	6,70	7,40	15,56
	Velikost okna 100 aa									
Nei-Gojobori	32,94	42,97	21,09	22,20	36,92	20,90	27,27	22,78	25,17	43,10
Li-Wu-Luo	47,83	61,12	49,04	48,29	68,18	43,88	48,49	47,62	51,57	48,32
Comeron	24,59	27,06	21,71	17,31	30,87	27,47	24,35	37,21	36,63	33,93
Shoda	22,22	26,67	17,55	10,68	27,45	16,01	16,22	21,21	23,15	33,93
	Velikost okna 200 aa									
Nei-Gojobori	62,52	68,90	50,36	52,17	61,84	37,50	52,02	40,37	45,91	76,72
Li-Wu-Luo	77,82	80,53	88,14	84,28	89,50	80,60	85,26	96,92	90,64	84,05
Comeron	36,04	32,23	40,35	36,57	50,33	26,49	38,76	51,04	43,01	47,23
Shoda	35,92	32,23	38,08	32,16	49,50	21,31	34,13	33,14	34,66	47,23

Na následujících grafech je zaneseno, kolik procent oken ze všech byly dle statistického z-testu vyhodnoceny s pozitivní či negativní selekcí. Pozitivní selekce měla u genu BRCA1 a vybraného datového setu malé procentuální zastoupení. Nejvíce oblastí s pozitivní selekcí se objevovalo u výpočtů mezi druhy *Homo sapiens* a *Mus musculus*. U okna velikosti 50 aa byly u této dvojice sekvencí největší procentuální hodnoty pozitivní selekce u všech tří metod (NG = 1,75%, LWL = 1,09%, COM = 2,74% se shodou v 1,04% oken). Při zvětšení okna na 100 aa, již metoda Li-Wu-Luo nezobrazila žádnou pozitivní selekci u této dvojice. Nýbrž se při této velikosti okna objevilo více oken s pozitivní selekcí u dvojice *Homo sapiens* a *Canis familiaris* (u okna 50 aa – NG = 0,82%, LWL = 0,00%, COM = 2,40%; u okna 100 aa – NG = 3,25%, LWL = 1,51%, COM = 3,75 se shodou v 1,51% oken). U okna velikosti 200 aa se objevila pozitivní selekce pouze opět u dvojice sekvencí *Homo sapiens* a *Canis familiaris*, ale pouze u metody Comeron – 0,36% oken.

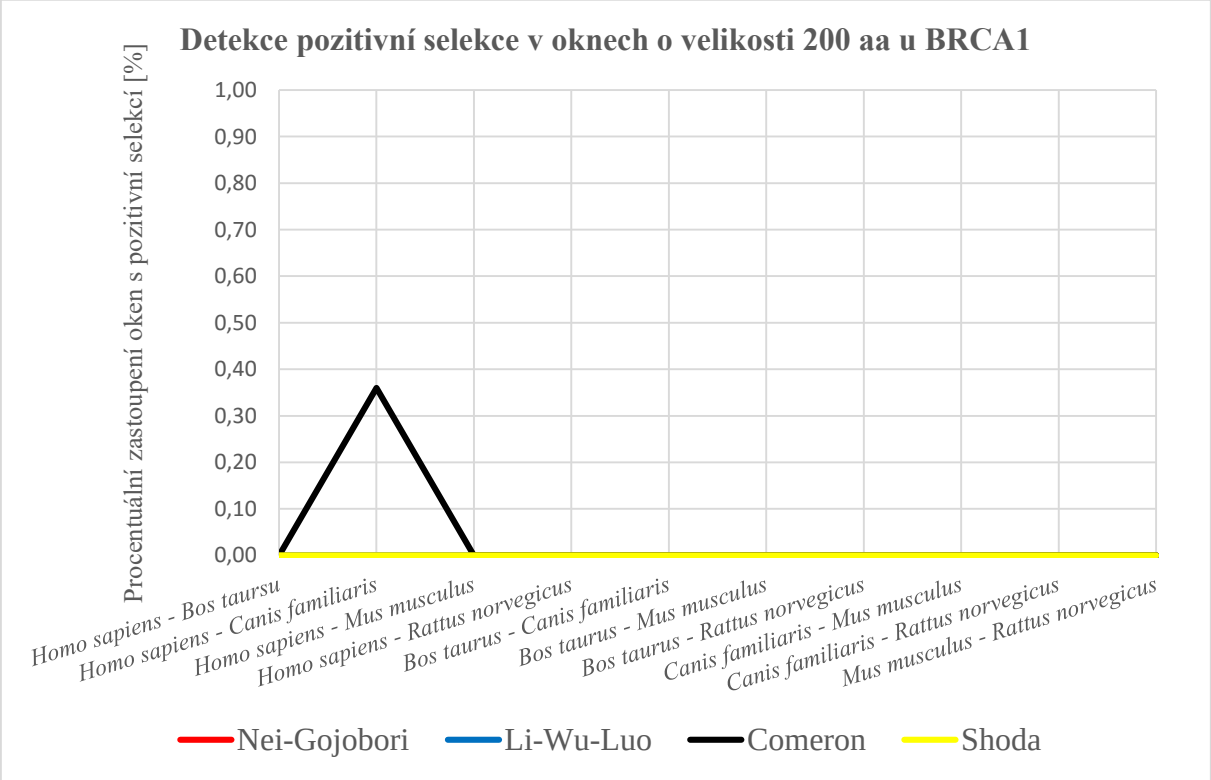
Více pozitivní selekce se objevují u metody Comeron, naopak nejméně oblastí s pozitivní selekcí určuje metoda Li-Wu-Luo. Shoda oken, ve kterých všechny metody detekovali pozitivní selekci, je téměř shodná s Li-Wu-Luo metodou. Dalo by se říct, že pokud byla pozitivní selekce určena metodou Li-Wu-Luo, ostatní metody taktéž tato okna označily za pozitivní selekci.



Obrázek 27: Detekce pozitivní selekce v oknech o velikosti 50 aa u BRCA1

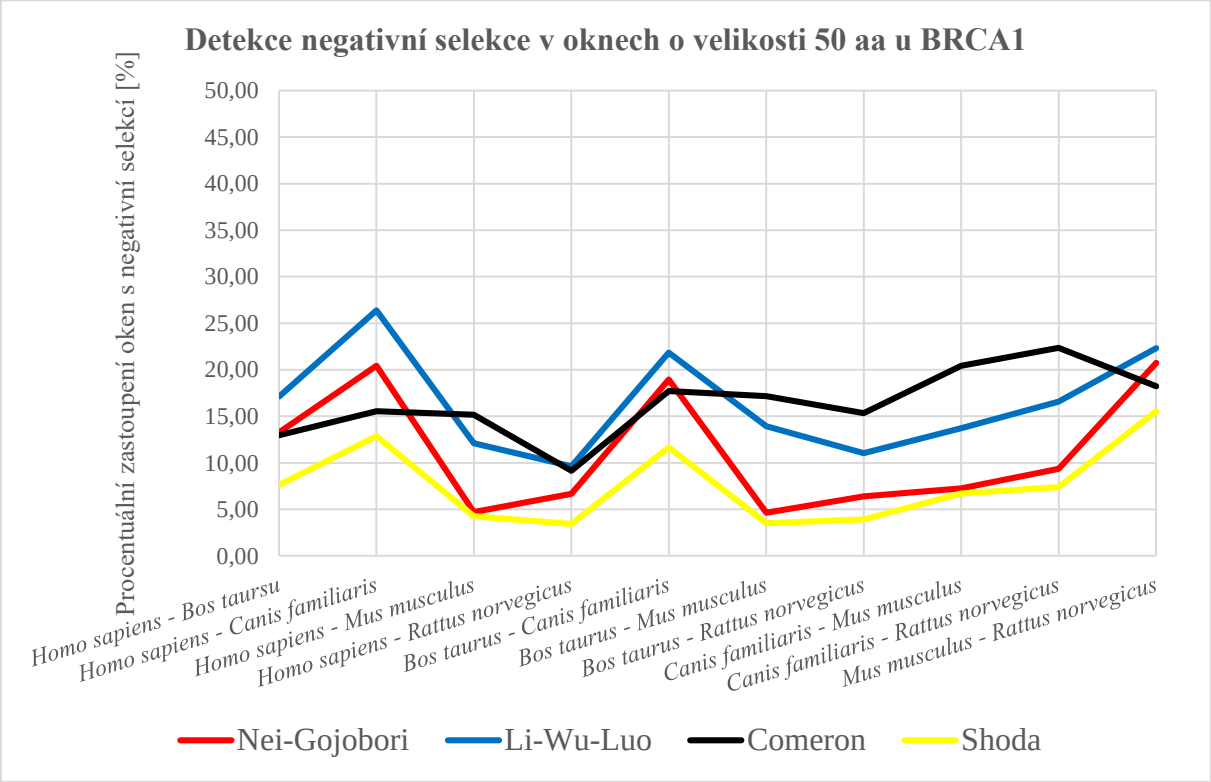


Obrázek 28: Detekce pozitivní selekce v oknech o velikosti 100 aa u BRCA1

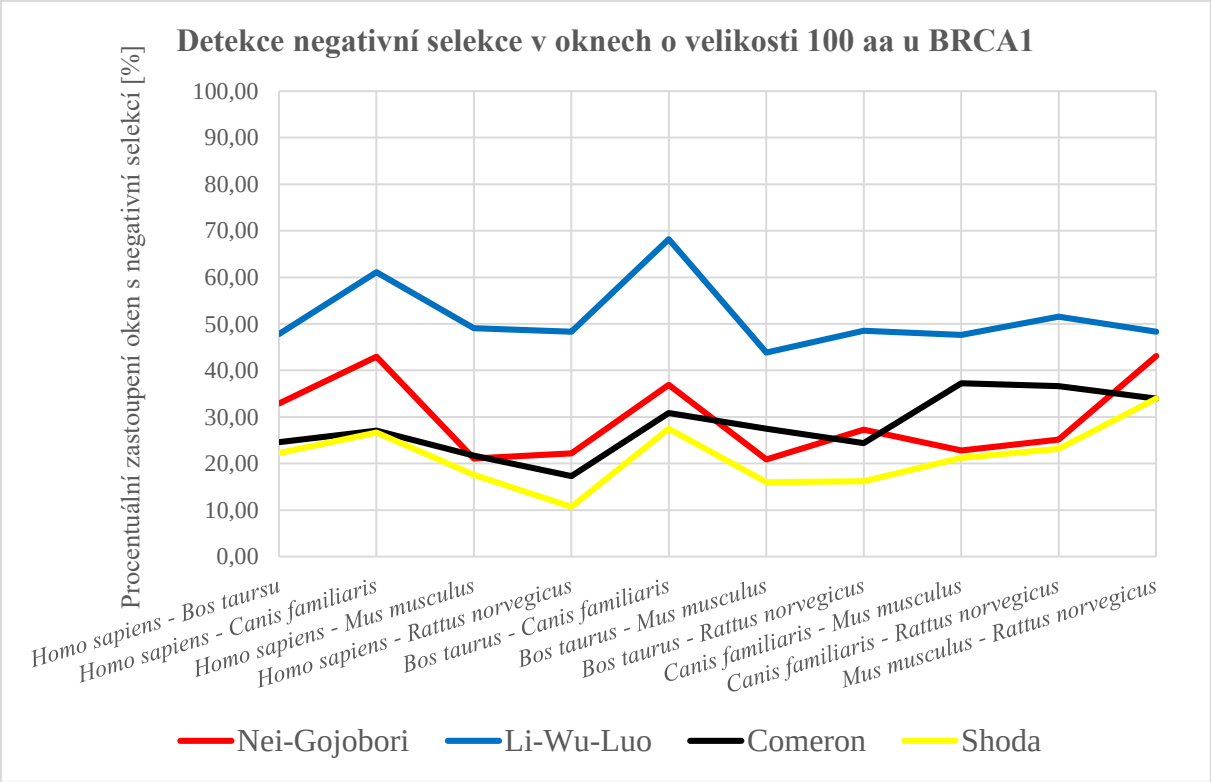


Obrázek 29: Detekce pozitivní selekce v oknech o velikosti 200 aa u BRCA1

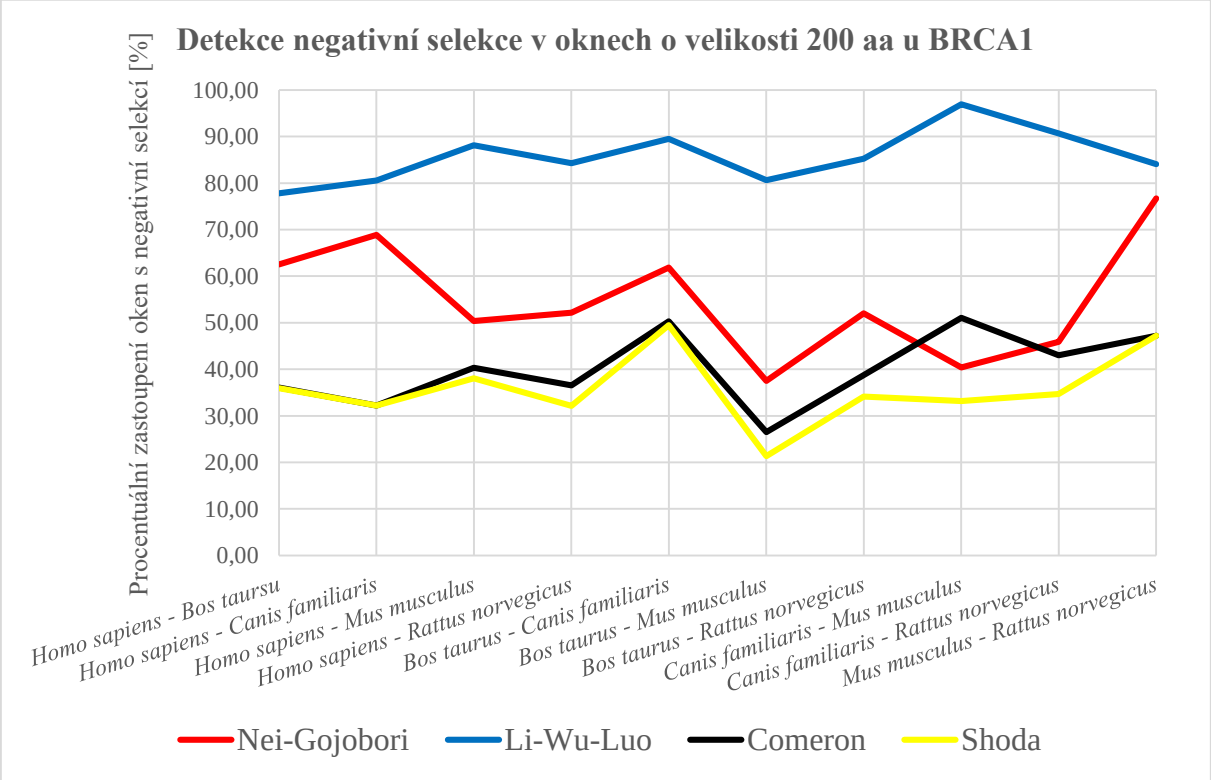
Při statistickém vyhodnocení negativní selekce jsme pozorovali, že roste procentuální zastoupení oken s potvrzenou negativní selekcí s velikostí okna. Metody Li-Wu-Luo a Nei-Gojobori detekují poměrově podobné oblasti u okna o velikosti 50 i 100 aa. Potvrzena negativní selekce u dvojice sekvencí *Homo sapiens* a *Canis familiaris* jsou dle okna zvětšující se – pro 50 aa okno: NG = 20,44% oken, LWL = 26,38% oken, COM = 15,53% oken se shodou 12,86% oken; pro okno 100 aa: NG = 42,97% oken, LWL = 61,12% oken, COM = 27,06% oken se shodou 26,67% oken; pro okno o velikosti 200 aa: NG = 68,90% oken, LWL = 80,53% oken, COM = 32,23% oken se stejnou shodou taktéž 32,23% oken. Hodnoty metody Comeron jsou u velikosti okna 50 a 100 aa bez závislosti na ostatní metody. U velikosti okna 200 aa je zřejmé, že výsledky metod Comeron a Li-Wu-Luo kopírují průběh, ale metoda Comeron statisticky určuje méně negativních selekcí než druhá metoda. Je možné, že je to způsobeno i možnou nepřesnou detekcí u této metody pro malé sekvence u malých oken. Na druhou stranu u velikosti okna 200 aa jsou menší výsledky oproti metodě Li-Wu-Luo způsobené rozšířením dvojitě-degenerativní skupiny. Například dvojice *Bos taurus* a *Canis familiaris* mají hodnoty výsledků testů – pro okno 50 aa: NG = 18,96% oken, LWL = 21,85% oken, COM = 17,71% oken se shodou 11,66% oken; pro okno 100 aa: NG = 36,92% oken, LWL = 68,18% oken, COM = 30,87% oken se shodou 27,45% oken; pro okno o velikosti 200 aa: NG = 61,84% oken, LWL = 89,50% oken, COM = 50,33% oken se shodou 49,50% oken.



Obrázek 30: Detekce negativní selekce v oknech o velikosti 50 aa u BRCA1



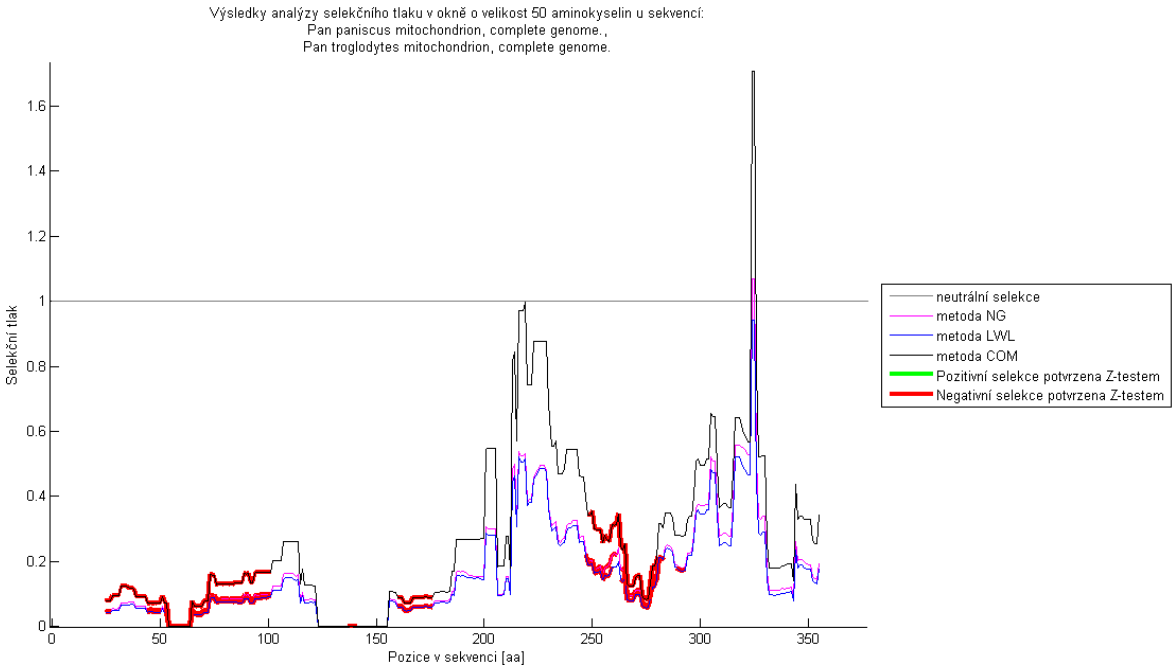
Obrázek 31: Detekce negativní selekce v oknech o velikosti 100 aa u BRCA1



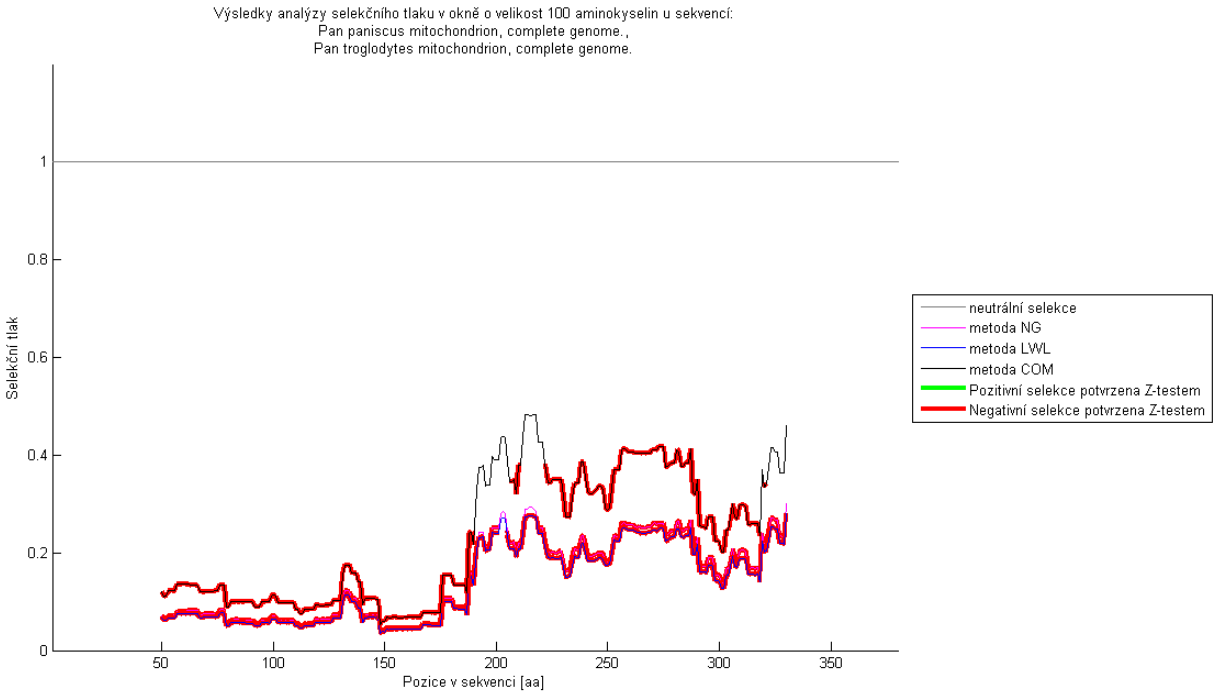
Obrázek 32: Detekce negativní selekce v oknech o velikosti 200 aa u BRCA1

U analýzy genu mtDNA pro cytochrom b z datového setu byly všechny hodnoty pozitivní selekce nulové. Mezi sekvencemi nedocházelo k upevňování pozitivních vlastností sekvence. Z toho důvodu se zde ani neobjevují tabulky či grafy pro tento typ selekce. Zaměříme se tedy pouze na popis vývoje negativní selekce u tohoto genu – viz grafy na Obrázek 40, Obrázek 41 a Obrázek 42 a hodnoty v Tabulka 18.

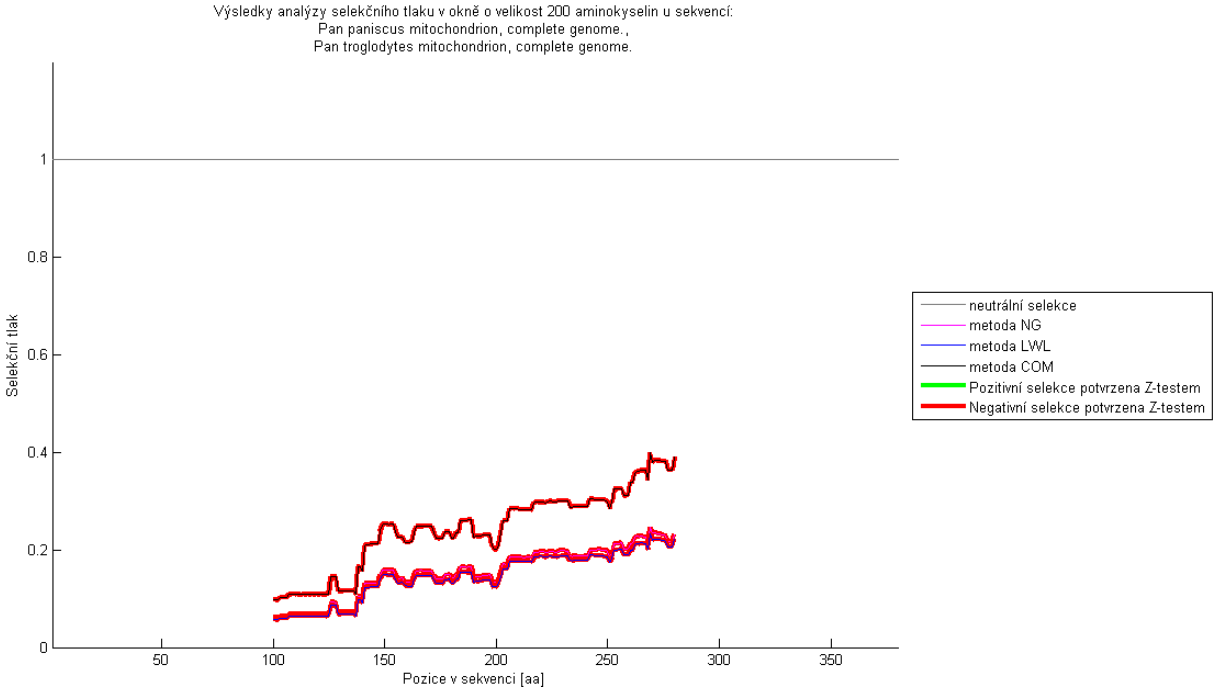
Opět můžeme i u tohoto genetického kódu, který odpovídá mitochondriálnímu kódu obratlovců, pozorovat, že s rostoucí velikostí okna se zvyšuje procentuální zastoupení oblastí statisticky potvrzených jako oblasti s negativní selekcí. Při okně 100 aa a 200 aa dosahují hodnoty 100%. Je to způsobeno tím, že velikosti sekvencí po zarovnání mají velikost 380 aa, které po kroku o jednu aminokyselinu prochází velké okno. Čtvrtina, respektive skoro polovina sekvence je oknem analyzovaná naráz. Pro představení jsou i zde zobrazeny průběhy selekčního tlaku pro jednu dvojici – *Pan paniscus* a *Pan troglodytes*, u které byla změna hodnot negativní selekce s různými okny největší – viz Obrázek 33, Obrázek 34 a Obrázek 35. Je zřejmé, že díky většímu oknu se průměrují informace ze sekvence, a z toho důvodu se celé sekvence posléze u velkého okna jeví jako 100% negativně selektivní. Tato dvojice u okna 50 aa vykazovala negativní selekci u metod NG = 27,49% oken, LWL = 32,02% oken, COM = 36,56% oken se shodou 26,86% oken, u okna o velikosti 100 aa byla negativní selekce stanovena u metod NG = 95,38% oken, LWL = 98,93% oken, COM = 87,54% oken se shodou 87,54% oken a u okna o velikosti 200 aa byla stanovena negativní selekce u všech metod 100% oken – viz Tabulka 18. Podobný vývoj byl i u ostatních dvojic mtDNA sekvencí pro gen cytochromu b z datového setu.



Obrázek 33: Výsledky selekčního tlaku mezi *Pan paniscus* a *Pan troglodytes* – gen cytochromu b z mtDNA, velikost okna 50 aa



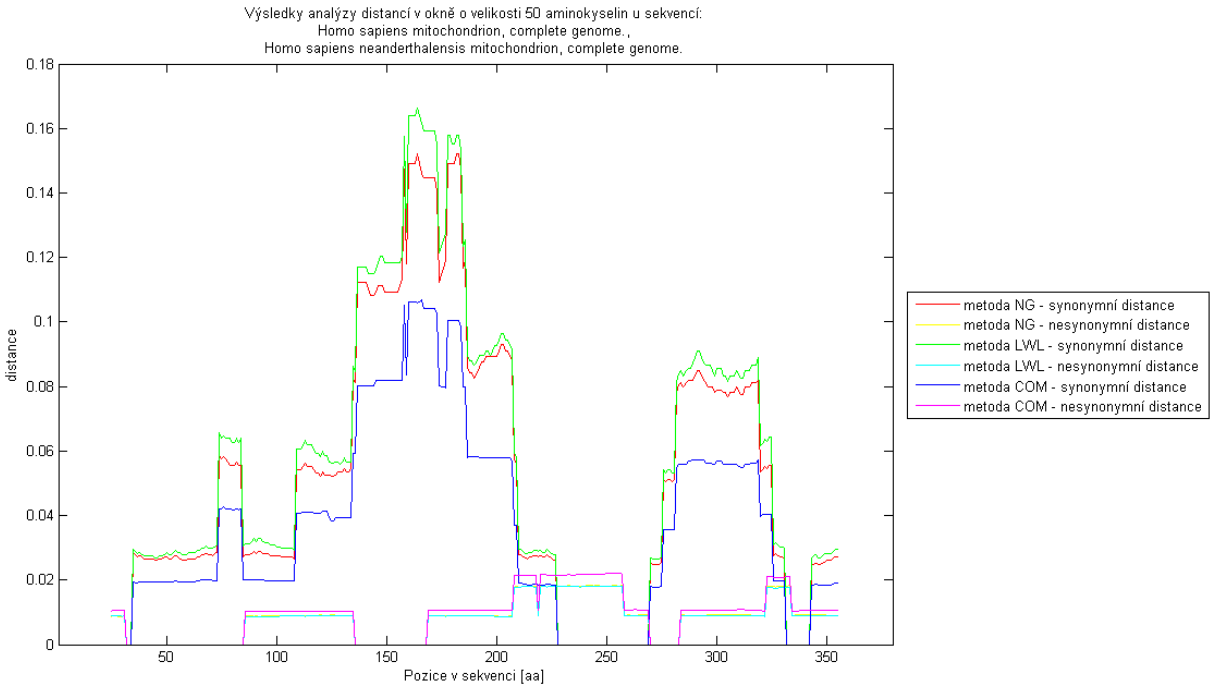
Obrázek 34: Výsledky selekčního tlaku mezi *Pan paniscus* a *Pan troglodytes* – gen cytochromu b z mtDNA, velikost okna 100 aa



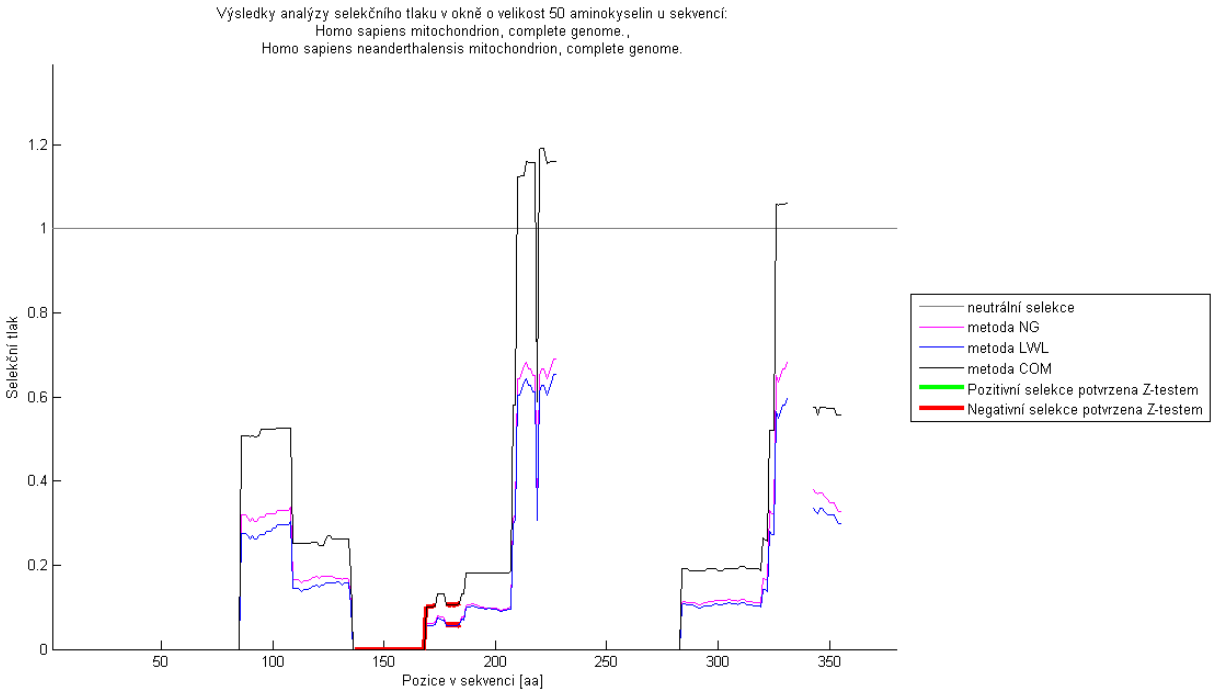
Obrázek 35: Výsledky selekčního tlaku mezi *Pan paniscus* a *Pan troglodytes* – gen cytochromu b z mtDNA, velikost okna 200 aa

Odlišnost je zřejmá u dvojice sekvencí *Homo sapiens* a *Homo sapiens neanderthalensis*. Jelikož se jedná o přímé předky, dalo by se předpokládat, že v průběhu evoluce došlo k pozitivní selekci, která ale použitými metodami v tomto případě nebyla vůbec potvrzena. Na druhou

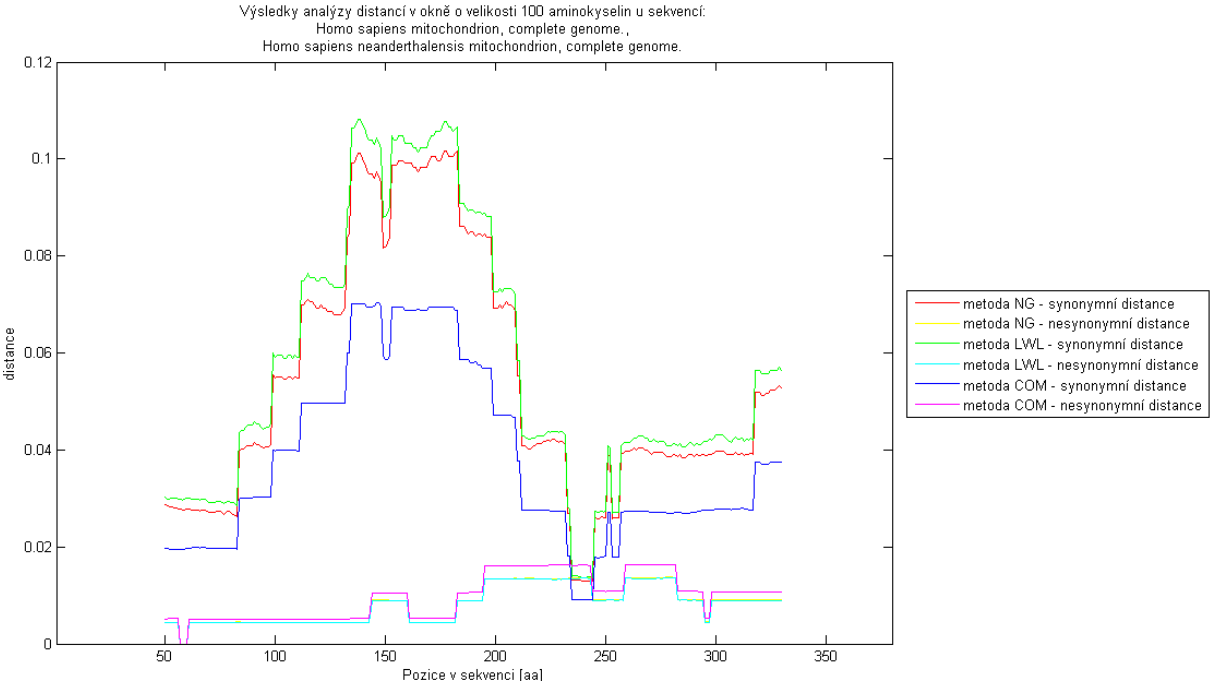
stranu negativní selekce v případě žádné zde prezentované analýzy a žádné velikosti okna nedosahovala 100%. Zbylé procento oken tedy podléhalo dle výsledků metod neutrální selekci. U okna 50 aa se ale u této dvojice nedá přesně interpretovat jejich výsledek, jelikož po zhlédnutí samotného průběhu výsledků selekčního tlaku a distancí – viz Obrázek 36 a Obrázek 37, lze pozorovat, že u několika úseků nebyl stanoven selekční tlak, jelikož hodnoty distancí odpovídaly nule u všech metod. Při zvětšení okna na 100 aa se již tyto úseky propočítaly díky zprůměrováním většího úseku sekvence – viz Obrázek 38 a Obrázek 39, ale jak již bylo zmíněno, díky tomu lze přijít o skrytou informaci ve velice krátkém úseku sekvencí.



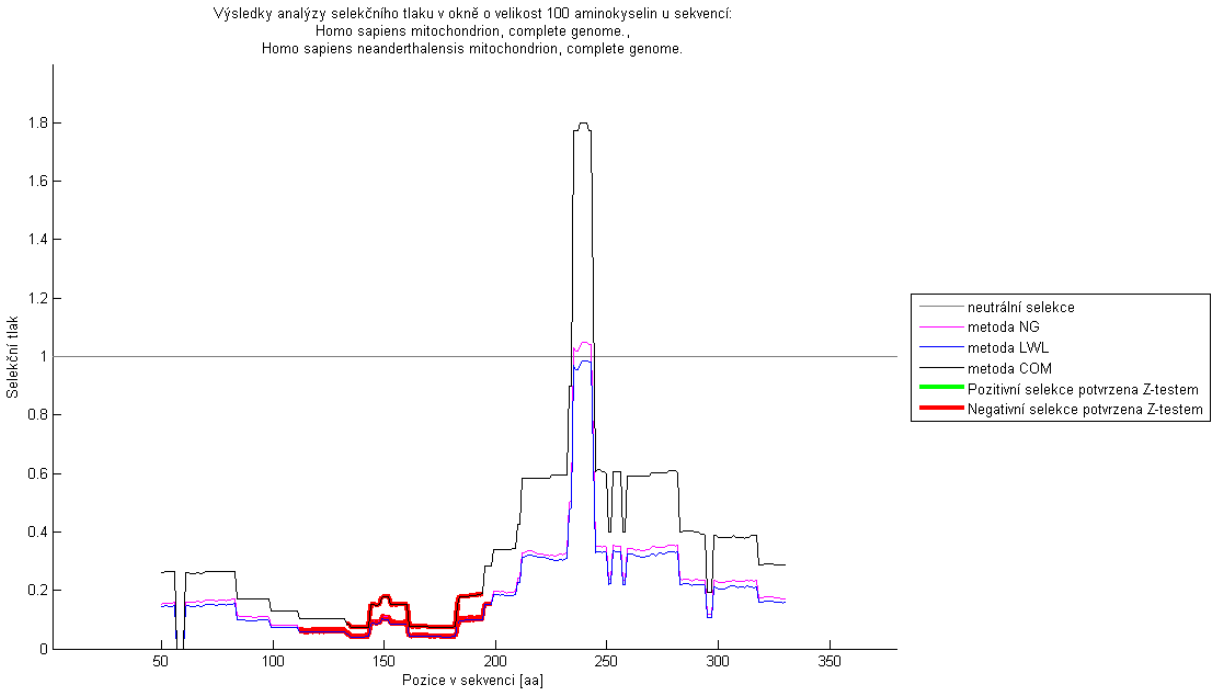
Obrázek 36: Výsledky synonymních a nesynonymních distancí s velikostí okna 50 aa mezi *Homo sapiens* a *Homo sapiens neanderhalensis* – gen cytochromu b z mtDNA



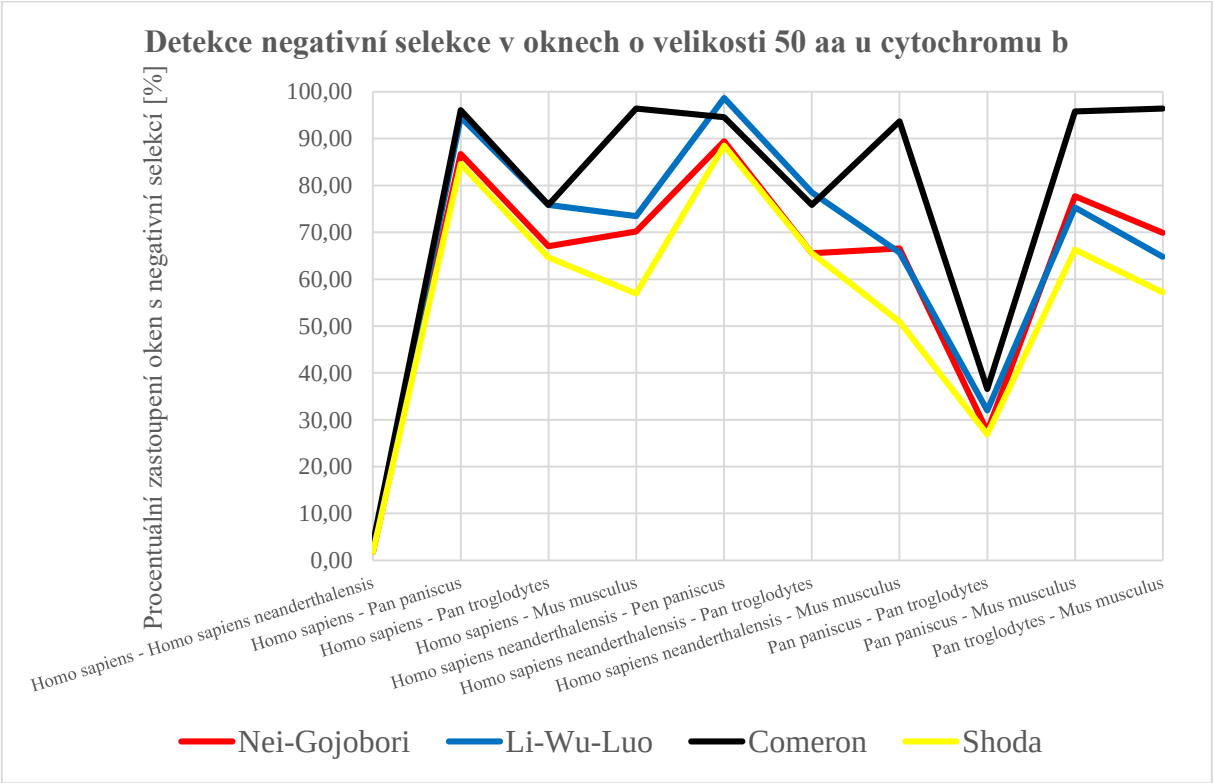
Obrázek 37: Výsledky selekčního tlaku s velikostí okna 50 aa mezi *Homo sapiens* a *Homo sapiens neanderhalensis* – gen cytochromu b z mtDNA



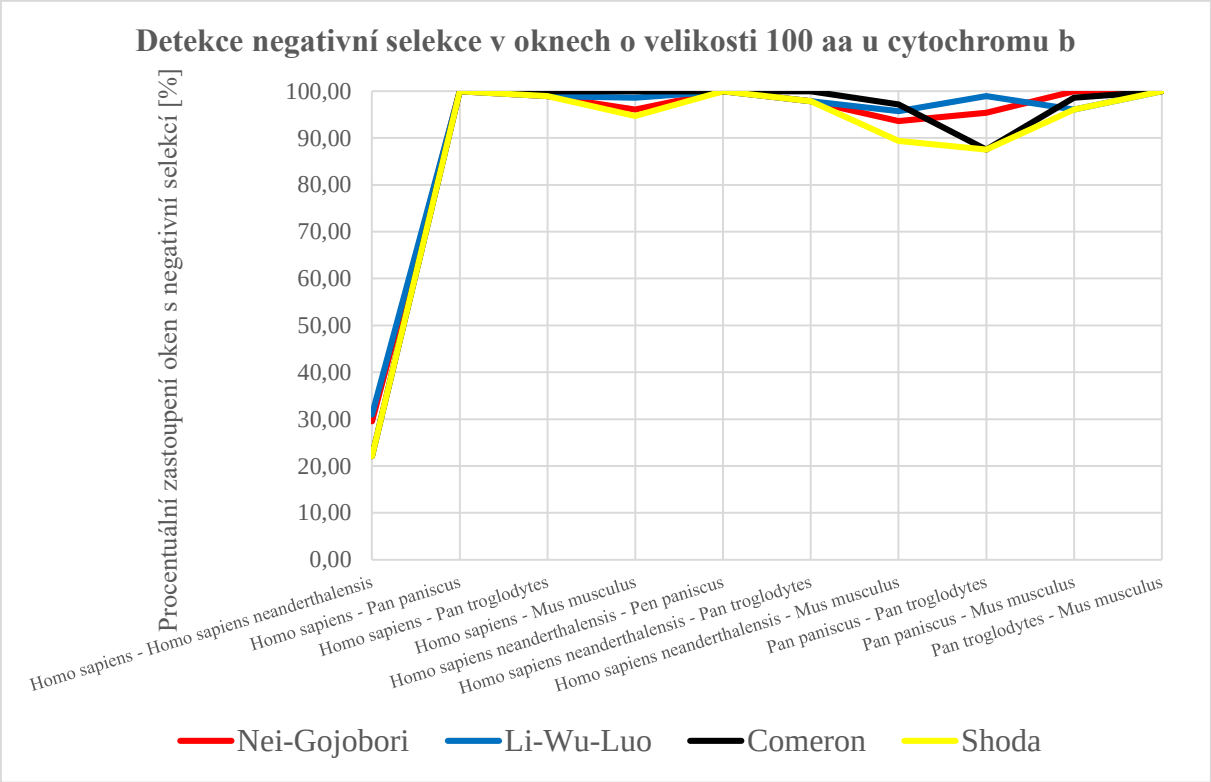
Obrázek 38: Výsledky synonymních a nesynonymních distancí s velikostí okna 100 aa mezi *Homo sapiens* a *Homo sapiens neanderhalensis* – gen cytochromu b z mtDNA



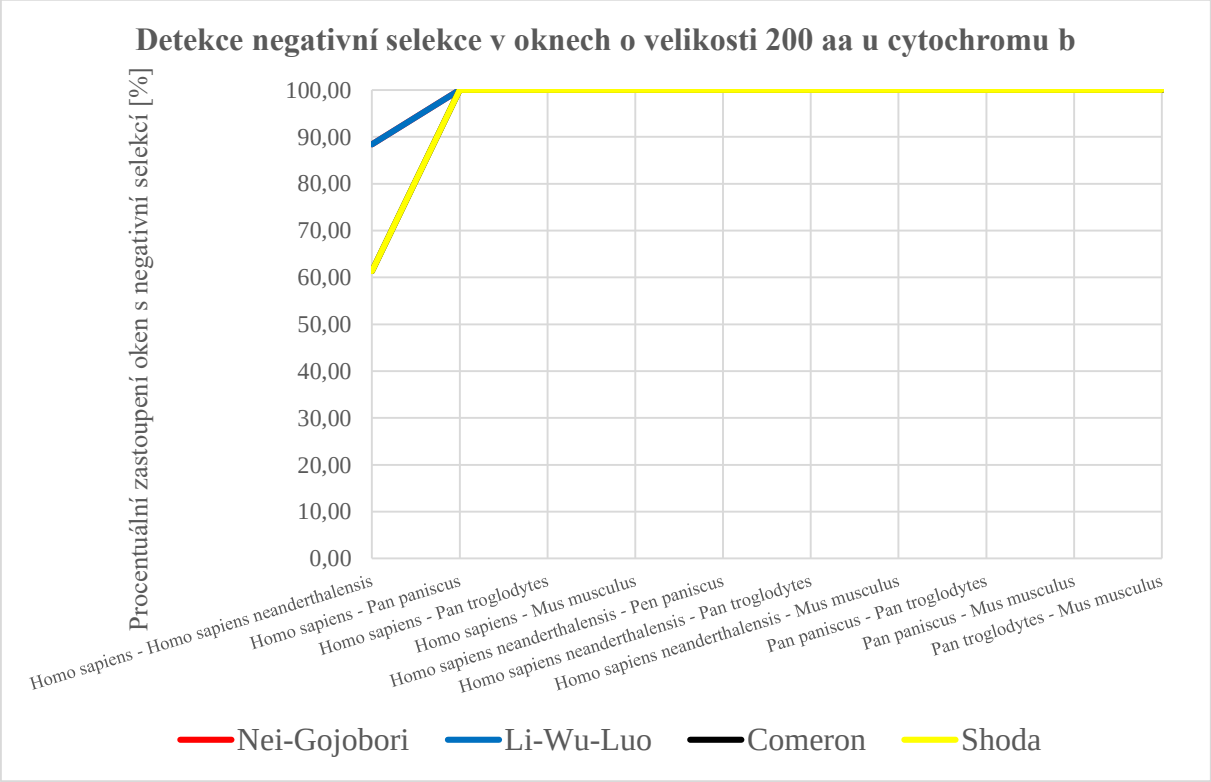
Obrázek 39: Výsledky selekčního tlaku s velikostí okna 100 aa mezi *Homo sapiens* a *Homo sapiens neanderhalensis* – gen cytochromu b z mtDNA



Obrázek 40: Detekce negativní selekce v oknech o velikosti 50 aa u Cytochromu b



Obrázek 41: Detekce negativní selekce v oknech o velikosti 100 aa u Cytochromu b



Obrázek 42: Detekce negativní selekce v oknech o velikosti 200 aa u Cytochromu b

Tabulka 18: Procentuální zastoupení oken s detekcí negativní selekce u datového setu mtDNA – Cytochrom b [%]

	Homo sapiens - Homo sapiens neanderthalensis	*Homo sapiens - Pan paniscus*	*Homo sapiens - Pan troglodytes*	*Homo sapiens - Mus musculus*	*Homo sapiens neanderthalensis - Pen paniscus*	*Homo sapiens neanderthalensis - Pan troglodytes*	*Homo sapiens neanderthalensis - Mus musculus*	*Pan paniscus - Pan troglodytes*	*Pan paniscus - Mus musculus*	*Pan troglodytes - Mus musculus*
	Velikost okna 50 aa									
Nei-Gojobori	1,81	86,71	67,07	70,18	89,43	65,56	66,57	27,49	77,71	69,88
Li-Wu-Luo	2,11	94,56	75,83	73,49	98,68	78,55	65,66	32,02	75,30	64,76
Comeron	3,62	96,07	75,83	96,39	94,56	75,83	93,67	36,56	95,78	96,39
Shoda	1,81	84,59	64,65	56,93	88,52	65,56	50,90	26,89	66,27	57,23
	Velikost okna 100 aa									
Nei-Gojobori	29,54	100,00	98,93	96,10	100,00	97,86	93,62	95,38	100,00	100,00
Li-Wu-Luo	30,96	100,00	98,93	98,60	100,00	97,86	95,74	98,93	96,10	100,00
Comeron	22,06	100,00	100,00	100,00	100,00	100,00	97,16	87,54	98,58	100,00
Shoda	22,06	100,00	98,93	94,68	100,00	97,86	89,36	87,54	96,10	100,00
	Velikost okna 200 aa									
Nei-Gojobori	88,40	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
Li-Wu-Luo	88,40	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
Comeron	61,33	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
Shoda	61,33	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

7. Závěr

Výstupem diplomové práce je software vytvoření v programovém prostředí MATLAB pro výpočet selekčního tlaku třemi různými metodami – Nei-Gojobori, Li-Wu-Luo a Comeron. Tyto metody až na poslední jsou již v bioinformatickém toolboxu zahrnuty, ale v tomto programu byly naprogramovány dle originálních článků autorů. Pro uživatelské využití je k softwaru jednodušší prostředí v GUI. Lze jednoduše nastavovat vstupní parametry – výběr metod, sekvencí z veřejných databází, dvojí typ zarovnání sekvencí, skórovací matice, genetický kód sekvencí, velikost okna i výřez sekvencí. Výstupem jsou hodnoty selekčního tlaku, grafy oblastí s hodnotami selekčního tlaku a zvýrazněných statisticky signifikantních oblastí určitého typu selekce. Pro lepší pochopení a analýzu výsledků lze zobrazit i distance synonymních a nesynonymních změn.

Pro vytvoření těchto algoritmů bylo zapotřebí vypracovat literární rešerši k dané tématice. Základem problematiky selekce jsou informace z molekulární evoluce. Především jsme se zaměřili na strukturu genů, popsali jsme si proteosyntézu a rozebrali jsme rozdíly mezi genetickými kódy, které ovlivňují výsledný překlad z nukleotidů na aminokyseliny. Pro vysvětlení synonymních a nesynonymních mutací jsme si vysvětlili samotné genové mutace.

Důležité bylo rozebrat si evoluci způsobenou výběrem, respektive právě selekcí. Bylo vysvětleno na matematickém základě, jaké hodnoty odpovídají negativní, pozitivní či neutrální selekci a co jednotlivé stavy znamenají.

Následovala část zabývající se samotnými matematickými modely výpočtu selekčního tlaku. Byly popsány tři metody výpočtu selekčního tlaku, které již byly zmíněny. Matematické výpočty byly demonstrovány na smyšlených jednoduchých příkladech pro jistotu pochopení. K matematickým základům metod se připojily ještě informace o matematických modelech evolučních distancí – Jukes-Cantor a Kimura, kterými jsou modely aproximovány. Ke všem matematickým krokům byly popsány i výpočty jejich variance, která je důležitá pro statistické vyhodnocení výsledků.

Samotná kapitola byla věnována právě statistickému vyhodnocení selekčního tlaku. Abychom statisticky potvrdili pozitivní či negativní selekci, byly tyto stavy zadány jako alternativní hypotézy Z-testu. Pomocí jeho výsledků jsme na hladině významnosti 5% mohli při oboustranném testu potvrzovat pozitivní či negativní selekci. Zmínili jsme se i o Fisherově exaktním testu, který ale nebyl v této práci využit.

Jednoduše jsme si představili hlavní funkce realizovaného softwaru na detekci selekce. Zajímavostí programu je přidružená databáze, která před prvním použitím již obsahuje předem vypočtené některé parametry pro všechny triplety všech používaných genetických kódů. Zbytek parametrů je při prvním výpočtu zapsán do databáze a již se nadále nikdy nepočítají, pouze se z databáze vyhledávají.

Nakonec jsme tento software využili na dvou datových setech – jeden se sekvencemi se standardním genetickým kódem a druhý se sekvencemi s mitochondriálním genetickým kódem obratlovců. Na některých výsledcích jsme si demonstrovali určitá úskalí, která při výpočtu selekčního tlaku můžou nastat. Vyzkoušeli jsme všechny metody selekce popsané v práci a zkoušeli jsme rozdílnost výsledků s nastavením různé velikosti okna. Při této zkoušce metod jsme vynesly graficky, kolik oken metody detekovali jako pozitivní nebo negativní selekci a porovnali metody vůči sobě. V kolika oknech byl jejich výsledek shodný.

Literatura

[1] ROSYPAL, Stanislav. Úvod do molekulární biologie. 3., inovované vyd. Brno: Stanislav Rosypal, 2002, 904-1200 s. ISBN 80-902562-4-4.

[2] NEI, Masatoshi a Sudhir KUMAR. Molecular evolution and phylogenetics. Oxford: Oxford University Press, 2000, xiv, 333 s. ISBN 0195135857.

[3] ŠÍPEK, A. jr.: Genetika - Biologie; Váš zdroj informací o genetice a biologii. online: [dostupné z http://www.genetika-biologie.cz; stav k 1. 12. 2015]

[4] ROSYPAL, Stanislav. Nový přehled biologie. 1. vyd. Praha: Scientia, 2003, xxii, 797 s. ISBN 978-80-86960-23-4.

[5] ALBERTS, Bruce. Molecular biology of the cell. 4th ed. New York: Garland Science, c2002, xxxiv, 1463 s. ISBN 0815332181.

[6] GRIFFITHS, Anthony J. An introduction to genetic analysis. 7th ed. New York: Freeman, c2000, xvii, 860 s. ISBN 0-7167-3520-2.

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

[8] YANG, Z. a J. P. BIELAWSKI. Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution, 12/1/ 2000, 15(12), 496-503.

[9] NEI, M. a T. GOJOBORI. Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions, Oxford: Oxford Univeresity Press, 1986, Molecular Biology and Evolution, 1/9/1986, 3,418-426.

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

[11] LI, W. H., WU C. I. And C. C. Luo. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol (1985) 2 (2): 150-174. online: [dostupné z http://mbe.oxfordjournals.org/content/2/2/150.long , stav k 1.12.2015]

[12] PERLER, R., A. EFSTRATIADIS, P. LOMEDICO, W. GILBERT, R. KLODNER a J. DODGSON. The evolution of genes: the chicken preproinsulin gene. Cell. Maryland Heights, 1980, 20(2): 555-566.

[13] ŠKUTKOVÁ, Helena. Modely evoluce sekvencí proteinů (Přednáška). ÚBMI FEKT VUT Brno, 2014.

[14] NACHMAN, M.W., 2006. Detecting selection at the molecular level. Evolutionary Genetics, Concepts and Case Studies, edited by C.W. Fox and J.B. Wolf. Oxford University Press, Oxford [dostupné z http://ib.berkeley.edu/labs/nachman/pdfs/ nachman_book_revised2A.pdf, stav k 2.12.2015]

[15] ISL1 Gene- GeneCards. *GeneCards* [online]. 2008 [cit. 2016-01-01]. Dostupné z: http://www.genecards.org/cgi-bin/carddisp.pl?gene=ISL1

[16] SAITOU, Naruya. Introduction to evolutionary genomics. New York: Springer-Verlag, 2013. Computational biology, 17. ISBN 1447153030.

[17] CHOUDHURI, Supratim. Bioinformatics for beginners: genes, genomes, molecular evolution, databases and analytical tools. Amsterdam: Elsevier, 2014. ISBN 978-0-12-410471-6.

[18] RAUSHER, Mark D. Handout: dN/dS ratios. In: Principles of Evolution [online]. Durham, USA: DUKE University, 2005 [cit. 2016-03-21]. Dostupné z: http://sites.biology.duke.edu/rausher/DNDS.pdf

[19] ŽUROVCOVÁ, Martina. Populační a evoluční genetika. In: Katedra genetiky PřF JU: Genetika [online]. Katedra genetiky PřF JU, České Budějovice: Jihočeská univerzita, 2008 [cit. 2016-03-21]. Dostupné z: http://kgn.umbr.cas.cz/prednasky/240%20Genetika/Lekce08-09handout.pdf

[20] HURST, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends in Genetics, 9/1/ 2002, 18(9), 486-487.

[21] COMERON, J. M. A Method for Estimating the Numbers of Synonymous and Nonsynonymous Substitutions per Site. Journal of Molecular Evolution. Dec 1995, 41(6): 1152-9

[22] ŠEDA, Ondřej. Druhy selekce. In: Wikiskripta [online]. 2011 [cit. 2016-05-04]. Dostupné z: http://www.wikiskripta.eu/index.php/Soubor:Druhyselekce.png

[23] Clarkwalker, G. D. & Weiller, G. F. The Structure of the Small Mitochondrial-DNA of Kluyveromyces Thermotolerans Is Likely to Reflect the Ancestral Gene Order in Fungi. J Mol Evol 38, 593–601 (1994)

[24] TEMPERLEY, R., R. RICHTER, S. DENNERLEIN, R. N. LIGHTOWLERS a Z. M. CHRZANOWSKA-LIGHTOWLERS. Hungry Codons Promote Frameshifting in Human Mitochondrial Ribosomes. Science. 2010, 327(5963), 301-301. DOI: 10.1126/science.1180674. ISSN 0036-8075. Dostupné také z: http://www.sciencemag.org/cgi/doi/10.1126/science.1180674

[25] OSAWA, Syozo, Takeshi OHAMA, Thomas H. JUKES, Kimitsuna WATANABE a Z. M. CHRZANOWSKA-LIGHTOWLERS. Evolution of the mitochondrial genetic code I. Origin of AGR serine and stop codons in metazoan mitochondria. Journal of Molecular Evolution. 1989, 29(3), 202-207. DOI: 10.1007/BF02100203. ISSN 0022-2844. Dostupné také z: http://link.springer.com/10.1007/BF02100203

[26] Bove, J. M. Molecular features of mollicutes. Clin. Infect. Dis. 17 (Suppl. 1), S10–S31 (1993)

[27] HOFFMAN, David C., Richard C. ANDERSON, Michelle L. DUBOIS, David M. PRESCOTT a Z. M. CHRZANOWSKA-LIGHTOWLERS. Macronuclear gene-sized molecules of hypotrichs. Nucleic Acids Research. 1995, 23(8), 1279-1283. DOI:

10.1093/nar/23.8.1279. ISSN 0305-1048. Dostupné také z: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/23.8.1279

[28] Keeling PJ, Doolittle WF. A non-canonical genetic code in an early diverging eukaryotic lineage. EMBO J 15: 2285–2290 (1996)

[29] JACOBS, Howard T., David J. ELLIOTT, Veerabhadracharya B. MATH, Andrew FARQUHARSON a Z. M. CHRZANOWSKA-LIGHTOWLERS. Nucleotide sequence and gene organization of sea urchin mitochondrial DNA. Journal of Molecular Biology. 1988, 202(2), 185-217. DOI: 10.1016/0022-2836(88)90452-4. ISSN 00222836. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/0022283688904524

[30] Golderer G, Dlaska M, Grobner P, Piendl W. TTG serves as an initiation codon for the ribosomal protein MvaS7 from the archaeon Methanococcus vannielii. J Bacteriol 177: 5994–5996 (1995)

[31] Santos MA, Keith G, Tuite MF. Non-standard translational events in Candida albicans mediated by an unusual seryl-tRNA with a 5′-CAG-3′ (leucine) anticodon. EMBO J 12: 607–616 (1993)

[32] S. Yokobori, T. Ueda, G. Feldmaier-Fuchs, S. Paabo, R. Ueshima, A. Kondow, K. Nishikawa, K. Watanabe Complete DNA sequence of the mitochondrial genome of the ascidianHalocynthia roretxi (Chordata, Urochordata) Genetics, 153 (1999), pp. 1851–1862

[33] TELFORD, M. J., E. A. HERNIOU, R. B. RUSSELL, D. T. J. LITTLEWOOD a Z. M. CHRZANOWSKA-LIGHTOWLERS. Changes in mitochondrial genetic codes as phylogenetic characters: Two examples from the flatworms. Proceedings of the National Academy of Sciences. 2000, 97(21), 11359-11364. DOI: 10.1073/pnas.97.21.11359. ISSN 0027-8424. Dostupné také z: http://www.pnas.org/cgi/doi/10.1073/pnas.97.21.11359

[34] HAYASHI-ISHIMARU, Yasuko, T. OHAMA, Yoshimi KAWATSU, a Syozo OSAWA. UAG is a sense codon in several chlorophycean mitochondria. Current Genetics. 1996-6-24, 30(1), 29-33. DOI: 10.1007/s002940050096. ISSN 0172-8083. Dostupné také z: http://link.springer.com/10.1007/s002940050096

[35] GAREY, James R., David R. WOLSTENHOLME, Yoshimi KAWATSU, Keiko NAKAMURA a Syozo OSAWA. Platyhelminth mitochondrial DNA: Evidence for early evolutionary origin of a tRNAserAGN that contains a dihydrouridine arm replacement loop, and of serine-specifying AGA and AGG codons. Journal of Molecular Evolution. 1989, 28(5), 374-387. DOI: 10.1007/BF02603072. ISSN 0022-2844. Dostupné také z: http://link.springer.com/10.1007/BF02603072

[36] NEDELCU, A. M., David R. WOLSTENHOLME, Yoshimi KAWATSU, Keiko NAKAMURA a Syozo OSAWA. The Complete Mitochondrial DNA Sequence of Scenedesmus obliquus Reflects an Intermediate Stage in the Evolution of the Green Algal

Mitochondrial Genome: Evidence for early evolutionary origin of a tRNAserAGN that contains a dihydrouridine arm replacement loop, and of serine-specifying AGA and AGG codons. Genome Research. 1989, 10(6), 819-831. DOI: 10.1101/gr.10.6.819. ISSN 10889051. Dostupné také z: http://www.genome.org/cgi/doi/10.1101/gr.10.6.819

[37] ELZANOWSKI, Andrzej a Jim OSTELL. The Genetic Codes. In: The National Center for Biotechnology Information [online]. Bethesda, Maryland, U.S.A., 2013 [cit. 2016-04-08]. Dostupné z: http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=t#SG1

[38] ADAMEC, Václav. Testy statistických hypotéz. Provozně ekonomická fakulta MENDELU, Zemědělská 1, Brno, 2005. Dostupné také z: http://user.mendelu.cz/urban/doc/gacr/genstat-testovani-hypotez.pdf

[39] DEONIER, Richard C., Simon TAVARÉ a Michael S. WATERMAN. Computational genome analysis: an introduction. New York: Springer, 2005. ISBN 03-879-8785-1.

[40] FAJMON, Břetislav, Irena HLAVIČKOVÁ a Michal NOVÁK. Matematika 3. Brno: Ústav matematiky FEKT VUT v Brně, 2014.

[41] NEEDLEMAN, Saul. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology [online]. 1970, 48(3), 443-453 [cit. 2016-05-15]. DOI: 10.1016/0022-2836(70)90057-4. ISSN 00222836.

[42] HENIKOFF, S. a J. G. HENIKOFF. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences. 1992, 89(22), 10915-10919. DOI: 10.1073/pnas.89.22.10915. ISSN 0027-8424. Dostupné také z: http://www.pnas.org/cgi/doi/10.1073/pnas.89.22.10915

[43] ALTSCHUL S. F., GISH W., MILLER W., MYERS E. W. a D. J. Lipman. Basic local alignment search tool. J. Mol. Biol. 1990 Oct 5;215(3):403–10.

[44] HOLČÍK, Jiří, KOMENDA, Martin (eds.) a kol. Matematická biologie: e-learningová učebnice [online]. 1. vydání. Brno: Masarykova univerzita, 2015. ISBN 978-80-210-8095-9. Dostupné také z: http://portal.matematickabiologie.cz/

[45] Brose MS, Smyrk TC, Weber B, aj. Genetic Basis of Cancer Syndromes. Kufe DW, Pollock RE, Weichselbaum RR, aj., Holland-Frei Cancer Medicine. 6th edition. Hamilton (ON): BC Decker; 2003.

[46] DOSTÁL, V. Commons.wikimedia.org : Mitochon-drial_DNA_cs.svg [online]. 2009-05-31 [cit. 2016-05-18]. Dostupný pod licencí Creative Commons na WWW: http://commons.wikimedia.org/wiki/File%3AMitochondrial_DNA_cs.svg

[47] TAANMAN, Jan-Willem, H. F. ROSENBERG a M. NEI. The mitochondrial genome: structure, transcription, translation and replication. Biochimica et Biophysica Acta (BBA) - Bioenergetics. 1999, 1410(2), 103-123. DOI: 10.1016/S0005-2728(98)00161-3. ISSN 00052728. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/S0005272898001613

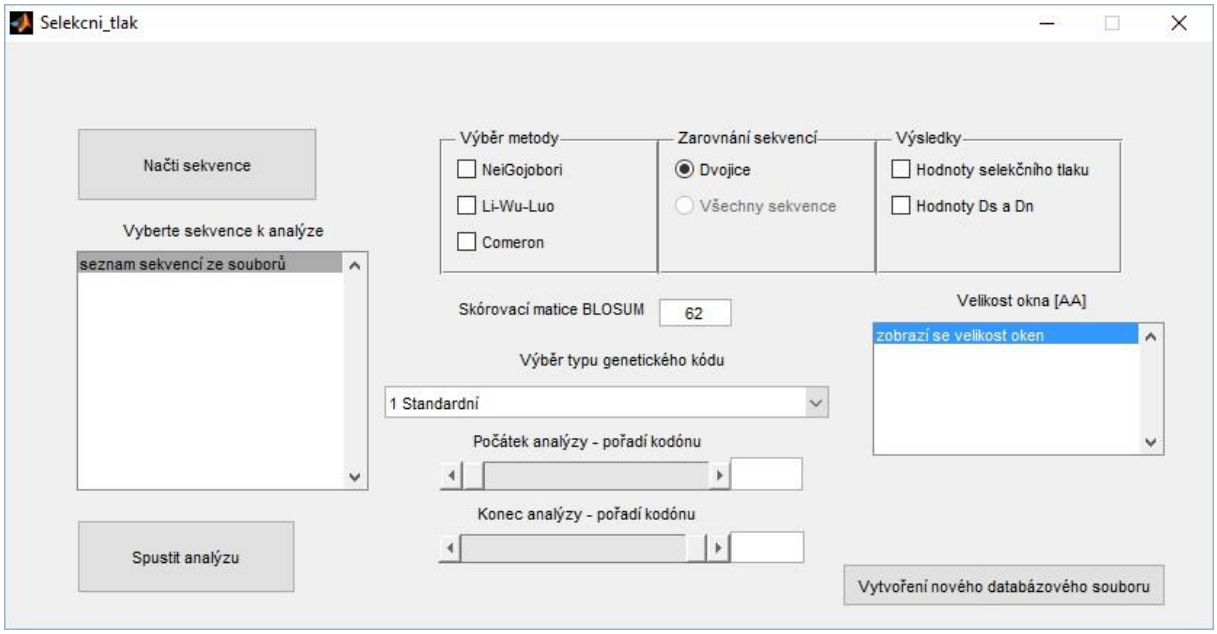
Seznam zkratek

aa	Aminokyselin
nt	Nukleotid
DNA	Deoxyribonukleová kyselina
mtDNA	Mitochondriální deoxyribonukleová kyselina
RNA	Ribonukleová kyselina
mRNA	Mediátorová ribonukleová kyselina
A	Adenin
C	Cytozin
G	Guanin
T	Thymin
U	Uracil
CDS	Kódující sekvence (z anglického coding sequence)
JC	Juckes-Cantor model
NG	Nei-Gojobori metoda
LWL	Li-Wu-Luo metoda
COM	Comeron metoda
BLOSUM	Typ skórovací matice (zkratka z anglického Blocks substitution matrix)
IUPAC	zkratka anglického názvu organizace International Union of Pure and Applied Chemistry
GUI	Uživatelské prostředí v programovém prostředí MATLAB (zkratka z anglického Grafical User Interface)

Přílohy

Příloha č. 1 – Manuál programu

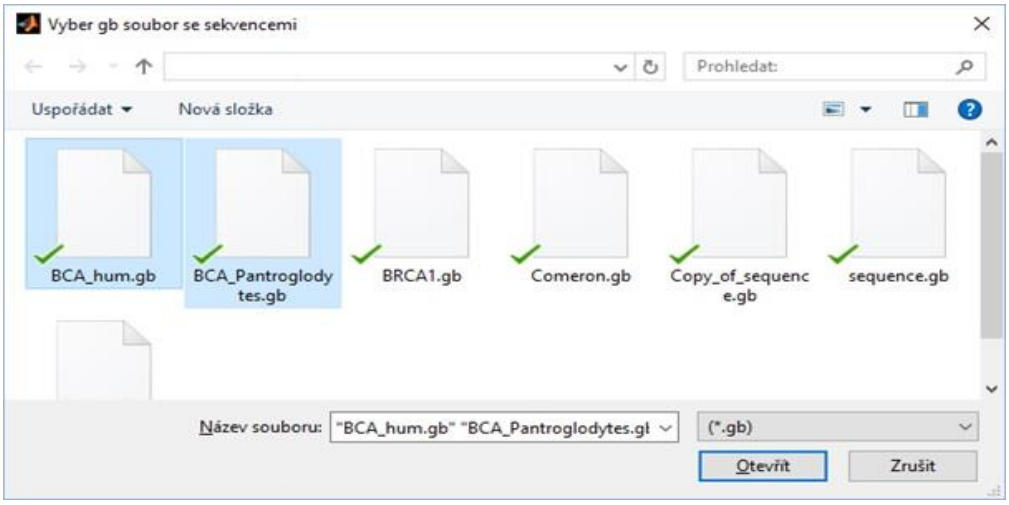
Program pro analýzu selekčního tlaku byl naprogramován v prostředí Matlab R2012b®. Součástí programu jsou soubory: `Slekcni_tlak.m` a `databaze.mat`. Program `Selekcni_tkal.m` obsahuje veškeré potřebné funkce pro vytvoření analýzy. MAT File `databaze.mat` obsahuje databáze výsledků proměnných pro samotnou analýzu a ukládají se do ní nově vypočtené výsledky parametrů, které jsou vypočteny v průběhu dalších analýz.



Obrázek 43: Uživatelské prostředí programu

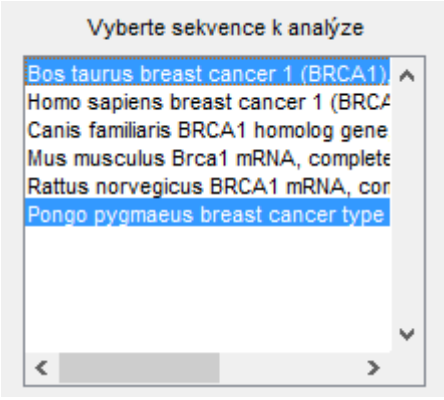
Pro samotné spuštění uživatelského prostředí GUI je nutné otevřít a následně spustit `Selekcni_tlak.m` v programu Matlab®.

Načíst sekvence je zapotřebí tlačítkem „Načti sekvence“ vlevo nahoře na Obrázek 43. Po kliknutí se objeví prohlížeč souborů – viz Obrázek 44, ve kterém si uživatel vybere soubory typu GenBank s koncovkou souboru *.gb. Je možné označit i více souborů najednou, pokud jsou staženy sekvence zvlášť. Upozorňuji, že předešlé nahrané sekvence budou z programu smazány.

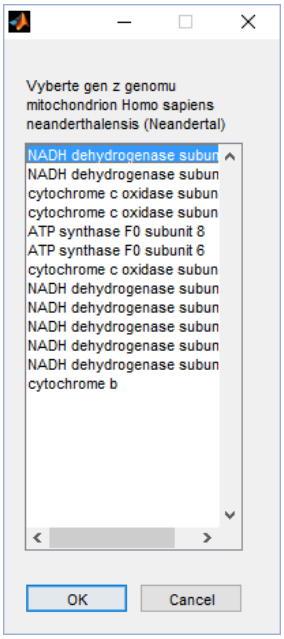


Obrázek 44: Prohlížeč souborů se soubory GenBank se sekvencemi

Dalším krokem ke správnému nastavení programu je výběr sekvencí, které byly nahrány. Uživatel si může ze seznamu názvů sekvencí z GenBank souborů – viz Obrázek 46 – vybrat dvě až všechny sekvence. Pro výběr více sekvencí lze použít tah myši nebo klikem myši s přidržením klávesy Ctrl. Modře označené sekvence budou vstupovat do analýzy. Pokud se nachází více genů v zadané sekvenci, program vyzve se seznamem všech genů, které v sekvenci nalezl, aby si uživatel vybral ten, který chce z dané sekvence analyzovat – viz Obrázek 45.

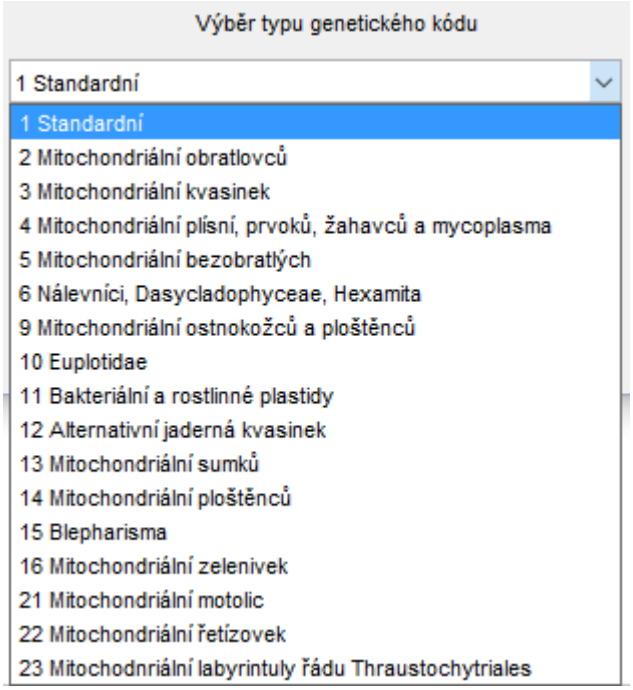


Obrázek 46: Výběr sekvencí k analýze



Obrázek 45: Výběr genu

Ke správnému výpočtu je zapotřebí vybrat, o jaký genetický kód se u sekvencí jedná. V rolovacím menu je na výběr ze sedmnácti různých genetických kódů, které jsou označeny užívanými čísly dle [37] a českým překladem. Pokud se nevybere správný genetický kód, může dojít ke špatným výsledkům. Tuto informaci si musí uživatel u každého souboru sekvencí zajistit sám.



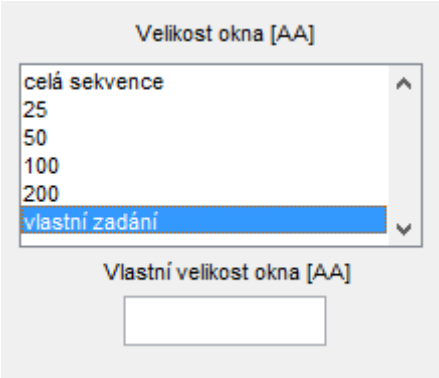
Obrázek 47: Výběr typu genetického kódu

Důležitým parametrem, který si uživatel může zatrhnout, je typ metody, kterou bude program počítat. Má na výběr z metod – Nei-Gojobori, Li-Wu-Luo a Comeron. Pro spuštění musí být vybrána alespoň jedna metoda. Mohou být vybrány všechny tři.

Při výběru více jak dvou sekvencí ze seznamu se zpřístupní možnost vybrat si, zdali uživatel chce sekvence zarovnávat globálním algoritmem Needleman-Wunsch každé dvě sekvence při výběru „Zarovnání sekvencí“ – „Dvojic“ nebo zdali si vybere „Všechny sekvence“ a chce zarovnat všechny sekvence algoritmem `multialigne`. K zarovnání se také pojí zadání hodnoty matice Blosum, která je přednastavena na defaultní hodnotu 62. Uživatel tuto hodnotu může přepsat v rozmezí hodnot 30 – 100, dle analyzovaných sekvencí.

Pro zobrazení žádoucího výsledku si uživatel v posledním odděleném bloku „Výsledky“ musí vybrat, jaké hodnoty chce zobrazit. Na výběr je hodnota pouze selekčního tlaku předem vybranými metodami nebo hodnoty distancí synonymních a nesynonymních substitucí. Obě tyto možnosti lze zobrazovat naráz.

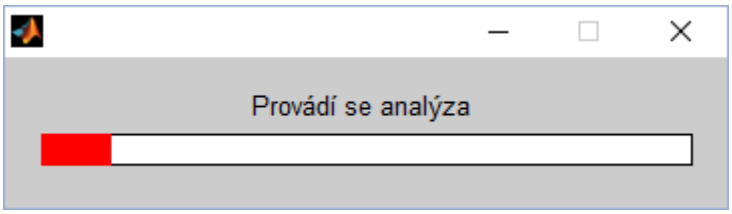
Další možnosti jsou volitelné. Dle potřeby si uživatel může vybrat z možnosti analýzy celé sekvence naráz nebo klouzavým oknem s variabilní velikostí počtu aminokyselin. Popřípadě lze vybrat velikost okna dle vlastní potřeby. Pro zobrazení dodatečného okna pro zapsání vlastní hodnoty velikost je zapotřebí, aby uživatel vybral „vlastní zadání“. Při volbě analýzy oknem je nutné, aby byly vybrány pouze dvě sekvence.



Obrázek 48: Výběr velikosti okna

Pro zpřesnění výsledku pouze na určitou oblast je možné využít oříznutí sekvencí dle zadaných hodnot prvního a posledního kodonu. Hodnoty lze zadat písemně do oken s čísly nebo pojízdným jezdcem.

K spuštění analýzy je již zapotřebí pouze spustit výpočet tlačítkem „Spustit analýzu“ vlevo dole. Program si ještě zkontroluje, zdali jsou zadané všechny vstupní parametry, které jsou k výpočtům zapotřebí. Pokud jsou veškeré zadané informace dostatečné, objeví se okno s grafickým zobrazením progresu analýzy. Po načtení se zobrazí výsledky, o které si uživatel požádal.



Obrázek 49: Grafické zobrazení progresu analýzy

Při spuštění program může upozornit na různé chyby. Kontroluje se, zdali jsou načteny sekvence nebo označené alespoň dvě sekvence k analýze. Upozorní uživatele chybovou hláškou na nevybrání žádné metody analýzy či žádaných výsledků. V takovém případě je nutné opravit zadané informace a zkusit spustit analýzu znovu.

Příloha č. 2 – Hodnoty distribuční funkce Φ(x) normované normální náhodné veličiny - výtah

Tabulka 19: Hodnoty distribuční funkce Φ(u), [40]

u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$
1,50	0,9331928	1,80	0,9640697	2,10	0,9821356	2,40	0,9918025	4,50	0,9999966
1,51	0,9344783	1,81	0,9648521	2,11	0,9825708	2,41	0,9920237	5,00	0,9999997
1,52	0,9357445	1,82	0,9656205	2,12	0,9829970	2,42	0,9922397	5,50	0,9999999
1,53	0,9369916	1,83	0,9663750	2,13	0,9834142	2,43	0,9924506		
1,54	0,9382198	1,84	0,9671159	2,14	0,9838226	2,44	0,9926564		
1,55	0,9394392	1,85	0,9678432	2,15	0,9842224	2,45	0,9928572		
1,56	0,9406201	1,86	0,9685572	2,16	0,9846137	2,46	0,9930531		
1,57	0,9417924	1,87	0,9692581	2,17	0,9849966	2,47	0,9932443		
1,58	0,9429466	1,88	0,9699460	2,18	0,9853713	2,48	0,9934309		
1,59	0,9440826	1,89	0,9706210	2,19	0,9857379	2,49	0,9936128		
1,60	0,9452007	1,90	0,9712834	2,20	0,9860966	2,50	0,9937903		
1,61	0,9463011	1,91	0,9719334	2,21	0,9864474	2,51	0,9939634		
1,62	0,9473839	1,92	0,9725711	2,22	0,9867906	2,52	0,9941323		
1,63	0,9484493	1,93	0,9731966	2,23	0,9871263	2,53	0,9942969		
1,64	0,9494974	1,94	0,9738102	2,24	0,9874545	2,54	0,9944574		
1,65	0,9505285	1,95	0,9744119	2,25	0,9877755	2,55	0,9946139		
1,66	0,9515428	1,96	0,9750021	2,26	0,9880894	2,56	0,9947664		
1,67	0,9525403	1,97	0,9755808	2,27	0,9883962	2,57	0,9949151		
1,68	0,9535213	1,98	0,9761482	2,28	0,9886962	2,58	0,9950600		
1,69	0,9544860	1,99	0,9767045	2,29	0,9889893	2,59	0,9952012		
1,70	0,9554345	2,00	0,9772499	2,30	0,9892759	2,60	0,9953388		
1,71	0,9563671	2,01	0,9777844	2,31	0,9895559	2,70	0,9965330		
1,72	0,9572838	2,02	0,9783083	2,32	0,9898296	2,80	0,9974449		
1,73	0,9581849	2,03	0,9788217	2,33	0,9900969	2,90	0,9981342		
1,74	0,9590705	2,04	0,9793248	2,34	0,9903581	3,00	0,9986501		
1,75	0,9599408	2,05	0,9798178	2,35	0,9906133	3,20	0,9993129		
1,76	0,9607961	2,06	0,9803007	2,36	0,9908625	3,40	0,9996631		
1,77	0,9616364	2,07	0,9807738	2,37	0,9911060	3,60	0,9998409		
1,78	0,9624620	2,08	0,9812372	2,38	0,9913437	3,80	0,9999277		
1,79	0,9632730	2,09	0,9816911	2,39	0,9915758	4,00	0,9999683		