Filozofická fakulta Univerzity Palackého v Olomouci

Katedra obecné lingvistiky

# Využití lingvistických metod v molekulární fylogenetice

*magisterská diplomová práce*

Autor: Bc. Hana Owsianková

Vedoucí práce: Mgr. Dan Faltýnek, Ph.D.

**Olomouc** **2017/2018**

## Prohlášení

Prohlašuji, že jsem magisterskou diplomovou práci „Využití lingvistických metod v molekulární fylogenetice“ vypracovala samostatně a uvedla jsem veškerou použitou literaturu a veškeré použité zdroje.

V dne Podpis

## Poděkování

Velmi děkuji Mgr. Danu Faltýnkovi, Ph.D. za jeho odborné vedení a podnětné nápady při konceptualizaci celé diplomové práce. Také děkuji Mgr. Vladimíru Matlachovi za pomoc s programováním, které jsem potřebovala pro realizaci výzkumu.

# Abstrakt

**Název práce**: Využití lingvistických metod v molekulární fylogenetice

**Autor práce**: Bc. Hana Owsianková

**Vedoucí práce**: Mgr. Dan Faltýnek, Ph.D.

**Počet stran**: 71

**Počet příloh**: 2

**Abstrakt**: Diplomová práce předkládá biosémiotické koncepty založené na analogii přirozeného jazyka a DNA, které vychází z chápání sekvencí DNA jako textu. Hlavním tématem je využití lingvistických metod v bioinformatice, předně pak pro účely molekulárně fylogenetických studií. Pro výzkumnou část je využita metoda Damerau-Levenshteinovy vzdálenosti a Bag-of-Words model. V prvním případě je testována hypotéza o příbuzenství druhu Malus domestica (jabloň domácí) a planého druhu Malus sieversii pocházejícího z oblasti Tyrkystánu v Centrální Asii. Druhý výzkum je zaměřen na mapování vztahů hybridů ke svým progenitorům tvořících tzv. U-model rodu Brassica (brukev). Pro oba výzkumy jsou použity genetické markery, konkrétně nukleární ribozomální marker ITS1 5.8S ITS2 a chloroplastový marker maturáza K. Cílem práce je prezentovat lingvistické metody jako vhodný nástroj pro zkoumání fylogenetických vztahů na mezidruhové úrovni.

**Klíčová slova**: biosemiotika, molekulární fylogenetika, lingvistické metody, Damerau-Levenshteinova vzdálenost, Bag-of-Words model, Malus, Brassica

# Abstract

**Title**: The utilization of linguistic methods in molecular phylogenetics

**Author**: Bc. Hana Owsianková

**Supervisor**: Mgr. Dan Faltýnek, Ph.D.

**Number of pages**: 71

**Number of appendices**: 2

**Abstract**: This diploma thesis presents biosemotic concepts based on the analogy of natural language and DNA, based on the understanding of DNA sequences as text. The main topic is the use of linguistic methods in bioinformatics, especially for molecular phylogenetic studies. For the research part, the Damerau-Levenshtein distance method and the Bag-of-Words model are used. In the first case, the hypothesis on the relation of Malus domestica (domestic apple tree) and the wild species Malus sieversii originating from the Tyrkystane area in Central Asia is tested. The second research is focused on the mapping of hybrids relations to their progenitors forming the so-called U-model of genus Brassica. Genetic markers, namely the nuclear ribosomal ITS1 5.8S ITS2 marker and the chloroplast marker maturase K, are used for both studies. The aim of this paper is to present linguistic methods as a suitable tool for the research of phylogenetic relationships at the interspecific level.

**Keywords**: biosemiotics, molecular phylogenetics, linguistic-like-tools, Damerau-Levenshtein distance, Bag-of-Words model, Malus, Brassica

# Obsah

# Úvod

*„Our own genomes carry the story of evolution, written in DNA, the language of molecular genetics, and the narrative is unmistakable.“*

*Kenneth R. Miller*

Molekulární fylogenetika je obor, který pro získání informací o evolučních vztazích organismů analyzuje molekulární znaky, jimiž jsou myšleny vybrané sekvence makromolekul – DNA, RNA a proteinů. Tento přístup začal v druhé polovině 20. století nahrazovat klasickou taxonomii vybudovanou Carlem Linné (Linnæi 1758), která je založena na morfologických znacích živých organismů, tzn. celkové vnější stavbě organismů – u rostlin se jedná např. o tvar, velikost a barvu plodu, květu nebo listu, velikost a tvar semen apod.; u zvířat se může jednat o celkovou stavbu kostry, pokrytí kůže a jeho zbravení (srst, peří, šupiny), typ larvy, tvar zobáku, rohů/parohů nebo kopyt aj. Molekulární fylogenetika pro své výzkumy využívá nejmodernější biochemické a bioinformatické metody, např. GISH/FISH (*genomic in situ hybridization/fluorescence in situ hybridization*), Multiple Sequence Alignment, Maximum Parsimony, Neighbour Joining, Maximum Likelihood ad. Jejím největším přínosem je revoluční změna v přístupu ke kategorizaci organismů do taxonů (především čeledí, tribů a rodů). Ačkoliv se tato vědecká disciplína neustále rozvíjí, mnoho otázek v oblasti genomiky a genetiky zatím ponechává nezodpovězených nebo na ně nedává jednoznačnou odpověď. To přináší výzvu jiným oborům, které mohou svou metodologií molekulárně fylogenetické výzkumy obohatit.

V této práci bych chtěla přiblížit koncept gramatik DNA, tedy paralel mezi přirozeným jazykem a „jazykem života“ – DNA. Věnuji se jeho vzniku, rozvoji, ale především praktickému využití v oblasti molekulární fylogenetiky. Pojetí zápisu genetických sekvencí jako textu se vyvíjelo od průlomového objevení struktury DNA v roce 1958 (F. Crick a J. Watson) a rozluštění genetického kódu v roce 1966 (M. Nirenberg, R. Holley, H. Khorana). Paralelně byly testovány možnosti využití lingvistických metod pro genomické a genetické studie. Své místo mezi standardními bioinformatickými metodami však začínají nalézat až v současnosti. Jejich potenciál není ani zdaleka vyčerpán a nemají standardizovanou podobu. Jako příklad efektivního využití lingvistických metod pro fylogenetický výzkum prezentuji Damerau-Levenshteinovu vzdálenost a Bag-of-Words model.

Úvodní teoretická část je věnována analogii přirozeného jazyka a DNA, jsou v ní prezentovány přístupy a koncepty hlavních představitelů biosémiotiky. Dále je pojednáváno o lingvistických metodách, které jsou v bioinformatice využívány, a o výhodách přístupu molekulární fylogenetiky oproti dřívějším taxonomickým postupům. Praktická část je rozdělena do dvou výzkumů. První se věnuje rodu Malus (jabloň) a testování hypotézy o vývoji Malus domestica (jabloně domácí) z planého druhu Malus sieversii pocházející z Centrální Asie z oblasti Tyrkystánu. Druhý výzkum je zaměřen na rod Brassica a využití lingvistických metod pro mapování vztahů mezi progenitory a jejich hybridy tvořící tzv. U-model. Pro oba výzkumy jsou využity molekulární markery, konkrétně nukleární ribozomální marker ITS1 5.8S ITS2 a chloroplastový marker maturáza K. Primárním cílem této studie je ukázat, že chápání genetických sekvencí jako textu a využití lingvistických metod pro jejich analýzu může velmi přispět k mapování fylogenetických vztahů a zaujmout místo vedle standardních bioinformatických postupů.

## Analogie přirozeného jazyka a DNA

*„The deciphering of the genetic code has revealed our possession of a language much older than hieroglyphics, a language as old as life itself, a language that is the most living language of all – even if its letters are invisible and its words are buried in the cells of our bodies."*

*George a Muriel Beadle, 1966*

Rozluštění genetického kódu v roce 1966 (M. Nirenberg, R. Holley, H. Khorana) bylo významným průlomem na poli genetiky a zároveň velkou inspirací k hledání paralel mezi genetickým kódem a přirozeným jazykem. Postupně se jazykové pojmy staly běžnou součástí terminologie molekulární biologie, např. kód, kódování, informace, přepis (transkripce), překlad (translace), řeč, zápis, editace (sestřih) ad. Jejich užívání se natolik ustálilo, že zcela přestaly být chápány jako metafory – *„Proteosyntéza – biosyntéza bílkovin – představuje pochod, při kterém se genetická informace uchovávaná a předávaná v řeči DNA exprimuje, tj. realizuje formou bílkovin. Proteosyntéza se také nazývá translace – překlad, neboť představuje biochemický převod informace z řeči nukleových kyselin do řeči bílkovin."* (Jonák 2007, s. 195)

Pro označení paralely mezi přirozeným jazykem a DNA nebo mezi živým tvarem a strukturami jazyka je užíván pojem *gramatiky DNA*. Toto téma spadá do oblasti zájmu biosémiotiky, disciplíny, která spojuje biologii a sémiotiku. Jejím hlavním záměrem je ukázat, že semióza je základní složkou života, tj. že ve všech živých organismech existují znaky a významy (Barbieri 2009). Termín „biosémiotika" poprvé použil Friedrich Salomon Rotschild (1962), k samotnému ustavení oboru došlo v návaznosti na Thomase Sebeoka a Jakoba von Uexkülla roku 2001 na mezinárodní biosémiotické konferenci *Gathering in Biosemiotics*. V důsledku odlišného chápání sémiotických konceptů (znak, signál, interpretace, kód nebo význam), se v rámci biosémiotiky vyčlenilo několik směrů (podle Kull 2007): fyzická biosémiotika (Howard Pattee), darwinistická biosémiotika (Howard Pattee, Terrence Deacon), zoosémiotika (Thomas Sebeok), znaková biosémiotika (Thomas Sebeok, Jessper Hoffmeyer), kódová biosémiotika (Marcello Barbieri) a hermeneutická biosémiotika (Anton Markoš). Analogií přirozeného jazyka a DNA se ve svých pracích zabývá především Roman Jakobson, Sungchul Ji, Anton Markoš, Edward Trifonov a Marcello Barbieri.

## Roman Jakobson

Roman Osipovič Jakobson (1896–1982), ruský lingvista a jeden ze zakladatelů Pražského lingvistického kroužku, je považován za jednoho z prvních, kdo konceptualizoval analogii mezi jazykem a DNA. K chápání DNA jako jazyka přispěl mnoha svými koncepty. Prvním z nich je paralela dvojí artikulace – tak jako lze v jazyce členit věty na slova, a ty dále na morfémy/fonémy, tak sekvence DNA lze dělit na triplety, a ty pak na jednotlivé nukleotidy. Roman Jakobson k sobě vztáhl analogické dvojice báze-foném/morfém, kodon-slovo a gen-věta. Abecedu DNA tedy metaforicky tvoří 4 báze (adenin, cytosin, guanin, thymin), které se kombinují do tripletů a kódují jednotlivé aminokyseliny. Jakobson dále připodobňuje řetězce DNA k linearitě v jazyce – ve výpovědi řadíme zvuky nebo písmena za sebou, stejně tak jsou v DNA kladeny za sebou báze. Pokračuje výkladem tří kodonů nekódujících aminokyselinu (stopkodon) jako interpunkčního znaménka. Dokonce i teorii binárních opozic, kterou původně vypracoval pro fonologický a morfologický systém, analogicky převedl na vztah vzájemně se doplňujících bází, jež v DNA vždy stojí proti sobě – cytosin-guanin, adenin-thymin/uracil (Jakobson 1971, s. 678–681).[1] Pro Jakobsona byl jazyk a život tím samým. Ve svých pojednáních (1973, 1974) o vztahu přirozeného jazyka a života obhájil oprávněnost užívání lingvistických termínů v genetice.

## Sungchul Ji

Sungchul Ji, povoláním původně farmakolog a toxikolog, se proslavil svým konceptem izomorfie mezi jazykem buněk, který nazývá *cellese*, a lidským jazykem, tzv. *humanese*. Inspiraci čerpal v biokybernetice a obecné molekulární teorii živých systémů (Ji 1991). Svou koncepci buněčného jazyka staví na předpokladu, že buňky v mnohobuněčných organismech spolu musí komunikovat za účelem svého přežití a vývoje, musí tedy mít svůj vlastní buněčný jazyk. Navíc jde ve své myšlence ještě o krok dál a naznačuje, že lidský jazyk je založen na buněčném jazyce. Samotný *cellese* definuje jako sebeorganizující se systém molekul, z nichž některé kódují, tzn. působí jako znaky nebo spouštěče pro ge-

[1] Nukleové báze (nuklotidy) jsou základní stavební jednotkou nukleových kyselin. Zozlišujeme báze purinové (adenin, guanin) a pyrimidinové (cytosin, uracil, thymin). Vždy vzniká vazba mezi jednou purinovou a jednou pyrimidinovou bází pomocí vodíkové vazby – guanin-cytosin, adenin-thymin/uracil.

nově řízené buněčné procesy. Pokud bychom Jiho pojetí vztáhli k Saussurově sémiotickému modelu, buněčný jazyk by užíval molekuly jako signifiant/označující a genově řízené buněčné procesy jako signifié/označované (Ji 1997, s. 17–19).

| | Human Language | Cell Language |
|---|---|---|
| 1. Alphabet (L) | Letters | 4 Nucleotides (or 20 amino acids) |
| 2. Lexicon (W) | Words | Structural genes (or polypeptides) |
| 3. Sentences (S) | Strings of words | Sets of genes expressed coordinately in space and time under the control of spatiotemporal genes[a] |
| 4. Grammar (G) | Rules of sentence formation | Laws of chemistry and physics of nucleic acids that determine the folding patterns of DNA according to nucleotide sequences and microenvironmental conditions. Only a small subset of grammatically folded (hence *syntactically* correct) chromatin structures is selected by evolution and hence carry genetic (i.e., *semantic*) information. |
| 5. Phonetics (P) | Physiologic structures and processes underlying phonation, audition, and interpretation | Conformational dynamics of DNA that enables the expression of genetic information through input of free energy via protein binding and/or ATP-dependent super coiling of DNA |
| 6. Semantics (M) | Meaning of words and sentences | Gene-directed cell processes driven by conformons[b] and intracellular dissipative structures (IDSs)[c] |
| 7. First Articulation | Formation of sentences from words | Organization of gene expression in space and time (through noncovalent interactions[d]) |
| 8. Second Articulation | Formation of words from letters | Organization of nucleotides (amino acids) into genes (polypeptides) (through *covalent interactions*[e]) |

**Obrázek 1**: Srovnání vlastností přirozeného a buněčného jazyka (humanese a cellese) dle S. Jiho (1999, s. 412)

Lidský i buněčný jazyk má podle Jiho šest základních společných rysů (abecedu, lexikon, věty, gramatika, fonetika, sémantika) a k nim přidává dvojí artikulaci (viz Obrázek1). Na základě těchto vlastností usuzuje, že DNA vyšších eukaryot (mnohobuněčných organismů) obsahuje dva druhy genů: *strukturální* – dle něj lexikální genetický kód (kódující DNA, u člověka 3%) a *spatiotemporální* (časoprostorově závislý) – tj. sémantický genetický kód (nekódující DNA, u člověka 97%). Syntaktický genetický kód pak prochází

napříč fyzickými i chemickými aspekty DNA každé molekuly (Ji 1999, s. 413). Ji chápe sekvenci DNA jako výraz komplexu všech procesů odehrávajících se v buňce.

## Edward Trifonov

Edward Nikolayevich Trifonov, ruský molekulární biofyzik a bioinformatik, podobně jako Ji, také zavedl pojem označující genetický jazyk – nazývá jej *gnomic*. Komunikace podle něj probíhá prostřednictvím nukleotidových sekvencí mezi molekulami v rámci procesů replikace, transkripce a translace.[2] Ve svých studiích (1989, 1990) se zabývá například lingvistickou komplexitou (diverzifikovanost, bohatost slovníku). Zatímco v textech přirozeného jazyka jsou dle něj jedním směrem postupně čteny všechny znaky, genetické texty jsou čteny několika různými způsoby – při užívání genetické informace jsou brány v potaz určité její znaky, zatímco jiné jsou opomíjeny; pokaždé dekódují jiné interakce a předpokládají overlapy (Trifonov hovoří o tzv. překrývajících se kódech; oproti Crickovi, který důsledně připomíná že genetický kód je nepřekrývající se - tzn. že je při využívání genetické informace pevně stanoven čtecí rámec tripletů.). Na rozdíl od lidských textů, které jsou tvořeny jediným kódem a nesou jednu informaci, genetické texty vytváří multikód, tj. nesou více informací a funkcí v jediné sekvenci. Trifonov se spolu s Popovou a Segalem (1996) věnují významu overlapů v genetických sekvencích a možnostem výpočtu strukturní složitosti libovolné lineární posloupnosti znaků (známých i nedefinovaných textů) na základě metody lingvistické komplexity. Konkrétně porovnávají proteinové sekvence s různými texty psanými v angličtině, italštině a velštině. Dokládají, že lidské texty jsou strukturně jednodušší než texty genetické, a to právě na základě výše popsaného rozdílného způsobu čtení.

[2] Replikace je proces, při kterém se genetická informace z jedné molekuly DNA přenáší do jiné molekuly. Tvorba bílkovin zahrnuje proces transkripce a translace. Transkripce představuje přepis genetické informace z DNA do mRNA (mediátorové RNA). Během translace je informace zapsaná v mRNA přenesena do primární struktury bílkovin, tzn. řetězce aminokyselin.

## Anton Markoš

Anton Markoš, profesí teoretický biolog, pokládá za společnou vlastnost lidských textů a „genetických textů“ (biologických makromolekul) to, že je možné je zaznamenat jako lineární zápisy prvků, které nesou nějakou informaci. Tyto prvky jsou určeny svým tvarem či vzhledem, pořadím a polohou. Pojmem informace zde odkazuje k Shannonově konceptu teorie informace, z nějž mnoho lingvistických metod užívaných v genetice vychází. Markoš bere sekvenci DNA jako text, který je možné rozčlenit na jednotlivá slova a poté zkoumat jejich frekvenci a distribuci – u jednoho nebo více organismů, jednotlivých genů/proteinů, celých genomů/proteomů (Markoš a kol. 2014, s. 58–84). Je považován za představitele tzv. hermeneutické biosémiotiky, klade důraz na způsob čtení genetického textu a jako jeden z možných uvádí proteosyntézu (proces tvorby bílkovin), kdy „mluvou jsou proteiny a jejich ‚syntax‘ a ‚sémantika‘ určují jak ‚výpověď‘, tj. fenotyp, tak výběr textů k dalšímu čtení“ (2003, s. 104).[3] Markoš zdůrazňuje, že pro fungování buňky není důležité pouze to, co je čteno, ale záleží také na způsobu, jakým je to čteno – tzn. informace není dána jen pořadím nukleotidů v DNA, ale také procesy interakce a interpretace v buňce.

## Marcello Barbieri

Marcello Barbieri, italský teoretický biolog, v roce 2008 začal vydávat časopis *Biosemiotics*, oficiální periodikum *International Society for Biosemiotic Studies*. Je zakladatelem disciplíny *Code Biology* (2015), kterou sám prezentuje jako studium všech kódů života pomocí standardních vědeckých metod, a jako odhalování dosud neprozkoumaného rozměru živého světa. Velkou pozornost věnuje organickým kódům (2006, 2008a), které díky evolučnímu zachování představují jediné neměnné entity, zatímco vše ostatní (tzn. fyzikální entity) se během vývoje mění – v souvislosti s jeho koncepcí organických kódů vyděluje např. genetický kód, metabolický kód, RNA kód, autoimunní kód ad. Podle Barbieriho je základní vlastností organických kódů arbitrárnost, tj. nemotivovanost výrazu a významu, což dokládá na vztahu mezi nukleovými kyselinami a proteiny, který je ustáleným způsobem (konvenčně) zprostředkován tzv. codemakery (např. ribozomy). Ostře se vymezuje (2008b) proti pojetí interpretace genetického textu v biosémiotické teorii

[3] Fenotyp jsou všechny pozorovatelné vlastností a znaky organismu. Jedná se o výsledek působení genotypu (genetická charakteristika jedince) a prostředí.

Antona Markoše. Svou kritiku staví na pojetí buňky jako biologického počítače, ve kterém dochází pouze k výrobní (*manufacturing*) a signalizační (*signalling*) semióze zprostředkované codemakery a nikoliv interpretací.

# Lingvistické metody v bioinformatice

*„Lingvistické metody nejsou schopny plně nahradit standardní srovnávací bioinformatické metody založené na přiřazování a hledání homologií, ale mohou posloužit jako doplňující nástroje k detailnější analýze. Navíc jsou výpočetně mnohem méně náročné."*

*Anton Markoš 2014, s. 84*

Bioinformatika je disciplína na pomezí informatiky a biologie, která se zabývá zpracováním a analýzou dat především v oblasti evolučních vztahů mezi živými organismy pomocí informačních technologií (http://bioinformatics.org). To, co dříve posuzovala taxonomie na základě morfologických znaků, tj. celkové vnější stavby organismu, dnes řeší molekulární fylogenetika pomocí standardních bioinformatických metod založených na analýze genetického textu (viz dále). Předmětem zájmu bioinformatiky je především mapování úseků DNA (hledání kódujících oblastí, nalezení primerů), získávání údajů o konkrétním genu či proteinu, hledání známých strukturních nebo sekvenčních motivů zkoumaného proteinu, porovnávání sekvencí dvou či více organismů a případné predikování funkčních rozdílů mezi nimi, posuzování míry příbuznosti genů v rámci větší genové skupiny (rod, tribe, čeleď) a následné usuzování evoluční historie a funkční rozrůzněnosti, nalézání funkčních genů v dosud neanotovaném úseku genomové sekvence atd. (Cvrčková 2006, s. 6).

Jednou ze základních metod, která umožňuje porovnávat a změřit podobnost genetických sekvencí (malá frakce aminokyselin či nukleotidů), je tzv. *sequence alignment* (přiřazení), jsou-li porovnávány více něž dvě sekvence, jedná se o *multiple sequence alignment*. V případě sekvencí s předpokládanou výraznou podobností, obsahujících dlouhé konzervativní úseky, je používáno globální přiřazení (*global alignment*), kdy jsou k sobě sekvence přiřazovány v celé své délce a jsou do nich zanášeny mezery (gaps) tam, kde jsou navzájem odlišné. Naopak pro evolučně vzdálenější sekvence je využíváno lokální přiřazení (*local alignment*), při němž jsou k sobě přiřazeny pouze jednoznačně shodné úseky (viz Obrázek 2). Váhy všech možných párů aminokyselin či nukleových bází určuje substituční matice (*scoring matrix, substitution matrix*). Pro sekvence nukleových bází se používá matice identity (*identity matrix* IUPAC), která všem párům i nepárům přiřazuje konstantní hodnotu, obvykle kladnou pro pár a zápornou/nulovou pro nepár. U proteinových sekvencí lze zvolit buď některou z matic PAM (*Point Accepted Mutation*) nebo

z matic BLOSUM (*BLOcks SUbstitution Matrix*), každá určuje pro jednotlivé páry a nepáry rozdílné penalizace (Cvrčková 2006, s. 28–45).


**Obrázek 2**: Rozdíl mezi lokálním a globálním přiřazením. (www.majordifferences.com)

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Ala** | 4 | | | | | | | | | | | | | | | | | | | |
| **Arg** | -1 | 5 | | | | | | | | | | | | | | | | | | |
| **Asn** | -2 | 0 | 6 | | | | | | | | | | | | | | | | | |
| **Asp** | -2 | -2 | 1 | 6 | | | | | | | | | | | | | | | | |
| **Cys** | 0 | -3 | -3 | -3 | 9 | | | | | | | | | | | | | | | |
| **Gln** | -1 | 1 | 0 | 0 | -3 | 5 | | | | | | | | | | | | | | |
| **Glu** | -1 | 0 | 0 | 2 | -4 | 2 | 5 | | | | | | | | | | | | | |
| **Gly** | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | | | | | | | | | | | | |
| **His** | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | | | | | | | | | | | |
| **Ile** | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | | | | | | | | | | |
| **Leu** | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | | | | | | | | | |
| **Lys** | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | | | | | | | | |
| **Met** | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | | | | | | | |
| **Phe** | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | | | | | | |
| **Pro** | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | | | | | |
| **Ser** | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | | | | |
| **Thr** | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | | | |
| **Trp** | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | | |
| **Tyr** | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | |
| **Val** | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

**Obrázek 3**: Příklad substituční matice – BLOSUM62 (wikipedia.org)

Spolu s jazykovými metaforami byly do genetiky vneseny lingvistické metody analýzy (označované jako *linguistic-like tools*). O jejich využití, především pro porovnávání sekvencí, pojednává Alexander Bolshoy (2003), významný teoretik v této oblasti. Konkrétně se věnuje metodě kontrastního slovníku, metodě kompozičního spektra a lingvistické komplexitě, které využívají rozkladu textu na n-gramy a vyhodnocují jeho diverzifikovanost. Mezi další průkopníky lingvistických metod v genomice patří například zmiňovaný Edward Trifonov. Ve spolupráci s ním vypracovali Michaela Zemková a Daniel Zahradník řadu analýz zaměřených na lingvistickou komlexitu a entropii genetických textů (viz Zemková 2016). Komplexitu genetických sekvencí zkoumal také Meeta Rani a Chanchal Mitra (1994). Kvalitativní strukturou genomů a proteomů, např. přítomností určitých „slov" napříč velkým množstvím proteinů, se zabýval Susumo Ohno (1992). Další uplatněnou lingvistickou metodou je testování platnosti tzv. Zipfova (mocninného) zákona na genetických sekvencích, podle nějž se slova v textu vyskytují s určitým statistickým rozložením (Mantegna 1995, Faltýnek – Matlach 2014).

Lingvistické přístupy k analýze genetických textů mohou sloužit jako alternativní metody vedle standardních metod užívaných v bioinformatice. Výše uvedené metody a výzkumy jsou jen několika málo příklady dynamicky se rozvíjejícího uplatnění teoretických poznatků biosémiotiky. Pro účely této práce se podrobněji zaměřím na metody, které byly z oblasti lingvistiky převzaty pro potřeby bioinformatických analýz nověji a zatím nebyly využity cíleně pro molekulárně fylogenetické studie – jedná se o Damerau-Levenshteinovu vzdálenost a Bag-of-Words model.

## Damerau-Levenshteinova vzdálenost

Damerau-Levenshteinova vzdálenost je pojmenována po průkopníkovi ve zpracování přirozeného jazyka (*Natural Language Processing*) Fredericku Damerau (1964) a informačním teoretikovi Vladimiru Levenshteinovi (1966). Formálně je vzdálenost definována jako počet úprav potřebných k transformaci jednoho řetězce na druhý, tyto řetězce mohou mít různou délku a jsou tvořeny konečnou abecedou. Transformacemi řetězce je zde myšlena *delece* (chybějící část), *inverze/substituce* (záměna částí), *inzerce* (začlenění části) nebo *transpozice* (prohození dvou sousedících částí). Původní koncept, který vytvořil Damerau, byl navržen jako editační nástroj pro identifikaci a opravu pravopisných chyb,

později byl uplatněn při sledování fonetických změn (Sanders et al. 2009) a také v oblasti lexikostatistiky v souvislosti s rekonstrukcí genetického stromu jazyků na základě analýzy výpůjček – viz Swadeshův / Leipzig-Jakarta seznam (Serva – Petroni 2007). Až spolupráce s Levenshteinem vedla k modifikaci výpočetního algoritmu (viz Obrázek 4) a využití metody v biologii k měření variace mezi genetickými sekvencemi.

$$d_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \\ d_{a,b}(i-2,j-2)+1 \end{cases} & \text{if } i,j > 1 \text{ and } a_i = b_{j-1} \text{ and } a_{i-1} = b_j \\ \min \begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

**Obrázek 4**: Algoritmus Damerau-Levenshteinovy vzdálenosti (wikipedia.org)

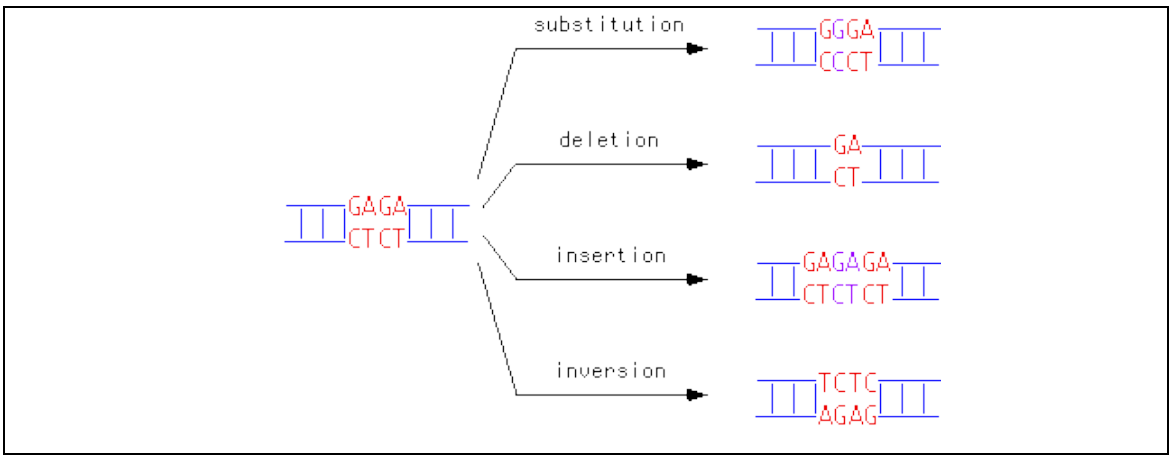
$d_{a,b}$ (i-1, j)+1 – odpovídá deleci

$d_{a,b}$ (i-1, j-1)+1($a_i \neq b_j$) – odpovídá substituci

$d_{a,b}$ (i, j-1)+1 – odpovídá inzerci

$d_{a,b}$ (i-2, j-2)+1 – odpovídá transpozici

V genetických sekvencích dochází běžně ke změnám jako je delece, inzerce, substituce a transpozice, ať už vlivem spontánních evolučních změn či hybridizace způsobené umělým zásahem člověka; jedná se o tzv. bodové mutace (*point genetic mutation*). Damerau-Levenshteinova vzdálenost je proto vhodnou metodou, která tyto změny mezi sekvencemi může detekovat (viz Obrázek 5). V praxi je podle výzkumného účelu z celého vzorku sekvencí vybrána jedna jako referenční a s ní jsou ostatní srovnávány, tzn. že je měřena vzdálenost všech sekvencí od zvolené referenční sekvence.




**Obrázek 5**: Typy bodových genetických mutací (carolguze.com)

Výsledná data jsou pak vizualizována, nejčastěji je užíván scatter plot (Schulz et al. 2012) a v případě sledování evolučních vztahů metoda hierarchického shlukování (Majorek et al. 2014). Uplatnění Damerau-Levenshteinovy vzdálenosti pro fylogenetické studie má také přesah do medicíny, příkladem je výzkum zaměřený na definování společného předka a evoluční trajektorie chronické lymfocytární leukémie (Sutton et al. 2014).

## Bag-of-Words model

Metoda Bag-of-Words spočívá v reprezentaci textu jeho slovy – výraz *Bag (taška, pytel)* vystihuje, že podstata modelu tkví v tom, zda se daná slova v dokumentu vyskytují, se zohledněním jejich frekvence, ale bez ohledu na jejich pořadí v textu. První zmínku o tomto modelu lze nalézt v článku amerického lingvisty a logického syntaktika Zelliga Harrise (1954, s. 156): *„... for language is not merely a bag of words but a tool with particular properties which have been fashioned in the course of its use.“* Jak je patrné, Harris považoval chápání jazyka jako „balíku slov“ za příliš redukující. Efektivní využití Bag-of-Words modelu pro zpracování, třídění a analýzu dat bylo rozvíjeno až od počátku 21. století (např. Toldo 2009, Zhang et al. 2010). V současnosti je metoda Bag-of-Words běžně využívána v oblasti zpracování přirozeného jazyka (*Natural Language Processing*), vyhledávání informací (*information retrieval*), počítačového vidění (*computer vision*) a klasifikace nejrůznějších dokumentů.

Intuitivně vnímáme, že jsou si dokumenty podobné, pokud obsahují stejná nebo podobná slova. Stejně tak umožňuje Bag-of-words model vyhodnocovat podobnost genetických sekvencí. Ačkoliv v lingvistice existuje několik přístupů k definování slova (významové kritérium, kritérium gramatičnosti, kritérium přemístitelnosti aj.), v *Natural Language Processing* je nejčastěji jako slovo chápána výrazová jednotka, která je z obou stran ohraničená mezerou. V případě genetických sekvencí nelze tato kritéria uplatnit, proto je využívána n-gramová analýza. Při ní jsou „genetické texty“ segmentovány na stejně dlouhé úseky, kombinace bází či aminokyselin, tzn. 2-gram = dvojkombinace bází/aminokyselin, 3-gram = trojkombinace bází/aminokyselin atd. (viz Obrázek 6). Jednotlivé texty jsou potom reprezentovány přítomností či nepřítomností těchto n-gramů (slov) a jejich frekvencí.

| **Typ sekvence** | **Jednotka** | **Příklad sekvence** | **1-gram sekvence** | **2-gramy sekvence** | **3-gramy sekvence** |
|---|---|---|---|---|---|
| **sekvence proteinu** | aminokyseliny | …Phe-Ser-Cys-Leu-His-Gly… | …Phe, Ser, Cys, Leu, His, Gly… | …Phe-Ser, Ser-Cys, Cys-Leu, Leu-His, His-Gly… | …Phe-Ser-Cys, Ser-Cys-Leu, Cys-Leu-His, Leu-His-Gly… |
| **sekvence DNA** | báze | …CGAT-GGAT… | …C, G, A, T, G, G, A, T… | …CG, GA, AT, TG, GG, GA, AT … | …CGA, GAT, ATG, TGG, GGA, GAT … |

**Obrázek 6**: Příklad n-gramové analýzy genetického textu

Analýza pomocí Bag-of-Words modelu je nejčastěji vizualizována pomocí histogramu, v případě sledování fylogenetických vztahů je pak vhodné hierarchické shlukování nebo MDS (*Multi Dimensional Scaling*). Jednoduchost a efektivita modelu vedla k jeho modifikaci pro nejrůznější účely – klasifikace různých textových či obrazových dokumentů, mapování a navigace (Filliat 2007), rozpoznávání výrazů ve tváři (Sikka et al. 2012) aj. Stejně jako u Damerau-Levenshteinovy vzdálenosti také zde najdeme využití v medicíně, například při interpetaci dat z oblasti biomedicínského inženýrství, jako jsou výsledky vyšetření EEG nebo EKG (Wang et al. 2013) či detekování infekce v organismu na základě počtu molekul T-receptoru (Lovato 2015).

## Molekulární fylogenetika

*„The species and the genus are always the work of nature [i.e. specially created]; the variety mostly that of circumstance; the class and the order are the work of nature and art.“*

*Carl von Linné, 1751*

Fylogenetika je obor systematické biologie, který se zabývá hledáním vývojových vztahů mezi organismy. Molekulární fylogenetika pro tyto účely používá molekulární znaky, tzn. pořadí monomerů v řetězcích biopolymerů, případně chemické a fyzikální vlastnosti biopolymerů.[4] Tato data původně sloužila pouze pro potřeby molekulární biologie, ale lze je využít také pro rozpoznávání jednotlivých druhů organismů, jejich třídění, pro zjišťování genealogické příbuznosti jedinců v rámci populace či druhu a pro rekonstrukci fylogeneze druhů nebo vyšších taxonů pomocí metod klasické či molekulární fylogenetiky.

Využití molekulárních znaků má oproti morfologickým několik výhod a specifik:

- Velikosti genomu organismů umožňuje pracovat s velkým množstvím molekulárních znaků.
- Molekulární znaky na úrovni nukleových bází, kterými se od sebe jednotlivé zkoumané druhy odlišují, na sobě nejsou obvykle závislé (někdy je však mezi molekulárními znaky vazba, která má funkční nebo historický původ).
- Molekulární znaky jsou vhodné také pro porovnávání organismů, které jsou si vzájemně nepříbuzné, a tudíž i nepodobné.
- Množství společně sdílených molekulárních znaků mezi dvěma druhy odráží míru příbuznosti těchto druhů, ale ne podobnost faktorů, které působily na jejich vývoj.
- Molekulární znaky jsou vhodným prostředkem pro studium *kladogeneze* (postupné odštěpování evolučních linií), avšak ne *anageneze* (postupné změny ve znacích jednotlivých evolučních linií (Flegr 2005, s. 439–442; podrobněji Page – Holmes 1998).

[4] Biopolymery (biologické makromolekuly) jsou polymery v živých organismech. Dělí se na polynukleotidy (DNA, RNA), polypeptidy (proteiny) a polysacharidy (karbohydráty, např. škrob, celulóza).

Jako první molekulární znaky, tzv. molekulární markery, byly využívaný proteiny zvané izoenzymy.[5] V současnosti existuje pro odhalovaní variability DNA několik typů markerů, jsou děleny do skupin podle způsobu jejich získávání: První skupinou jsou markery získané analýzou fragmentů DNA pomocí polymerázové řetězové reakce (PCR), tj. metodě rychlé a snadné replikace bází určitého úseku DNA – AFLP (*Arbitrary Fragments Length Polymorphism*), RAPD (*Random Amplified Polymorphic DNA*), ISSRs (*Inter Simple Sequence Repeats*), SCP (*Single Strain Conformation Polymorphism*), PCR (*Polymerase Chain Reaction*) a tzv. mikrosatelikty (*SSR – Simple Sequence Repeats, STR – Short Tandem Repeats*). Dalším typem jsou markery označované RFLP (*Restriction Fragment Length Polymorphism*), známé také jako restrikční fingerprinting, ty využívají k charakterizaci DNA přítomnost polymorfismu v délkách restrikčních fragmentů.[6] Třetí skupinou jsou celogenomové markery založené na sekvenování DNA, z nich je nejvyužívanější SNP marker (*Single Nucleotide Polymorphism*) označující variaci v jediném nukleotidu, která se vyskytuje v určité pozici v genomu (Flegr 2005, s. 42–52).

V praktické části této práce se zaměřím na analýzu fylogenetických vztahů v rámci rodu Malus (jabloň) a rodu Brassica (brukev) na základě lingvistických metod. Pro molekulárně fylogenetické studie je vždy lepší využívat více genetických markerů, aby použitá vizualizace dat (dendrogram, MDS) poskytla co nejpřesnější rekonstrukci fylogenetických vztahů. U rostlin jsou převážně kombinovány markery chloroplastového genu a ITS oblast (*Internal Transribed Spacer*) jaderné ribozomální DNA (např. Eriksson et al. 2003). ITS úsek nekódující DNA, který se nalézá mezi malou (SSU) a velkou (LSU) podjednotkou ribosomální DNA, a chloroplastové markery jsou často využívány ve fylogenetických studiích rostlin pro svou velkou variabilitu i mezi blízce příbuznými druhy. Ze stejného důvodu jsem se rozhodla tyto markery použít pro analýzy také v této práci.

[5] Izoenzym je enzym, který má několik chemických variant. Tyto varianty většinou vznikají v různých orgánech nebo tkáních, ale vždy plní podobnou funkci.

[6] Restrikční fragmenty jsou úseky DNA vzniklé jeho štěpením na specifických místech pomocí enzymu restrikční endonukleáza. Tento způsob štěpení se používá především při metodě zvané „polymorfismus délky restrikčních fragmentů“ (*restriction fragment length polymorphism).*

## Fylogeneze Malus domestica

*„Jablko nepadá daleko od stromu.“* *– české přísloví*

Rod Malus (jabloň) z botanického hlediska patří do řádu růžokvětých (Rosales), čeledě růžovitých (Rosaceae), podčeledě jabloňovitých (Maloideae) a tribu Maleae. Vzhledem k výraznému polymorfismu mnoha druhů, které jsou rozšířené v klimaticky rozmanitých oblastech, v současnosti neexistuje jednotná taxonomie rodu Malus. Odborníci se dokonce mnohdy neshodnou na určení a zařazení některých základních druhů (pro srovnání např. Koidzumi 1934, Rehder 1949, Blažek 1998, Brickell 2003). V závislosti na odlišné klasifikaci původních druhů, variet a hybridů se celkový počet pohybuje v rozmezí 30–55 druhů – např. server The Plant List (www.theplantlist.org), databáze všech známých rostlin, uvádí 35 druhů.

Rod Malus je také dělen na 7„sekcí“ (section – taxonomická jednotka mezi druhem a podrodem; dle Phipps et al. 1990):

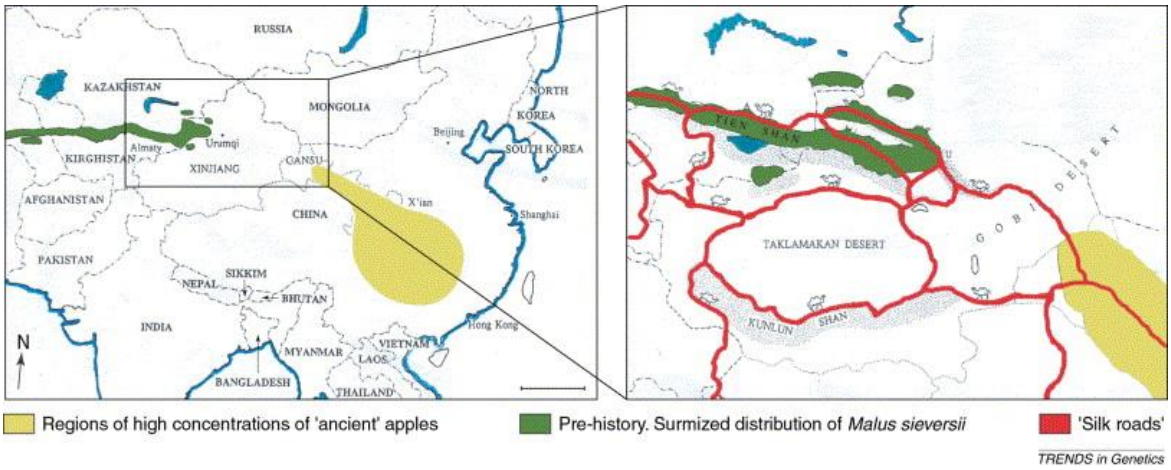
- **Malus sect. Chloromeles –** *M. angustifolia* (jabloň úzkolistá), *M. coronaria* (jabloň korunná), *M. ioensis* (jabloň zelenoplodá)
- **Malus sect. Docyniopsis** – *M. doumeri*, *M. tschonoskii* (jabloň Tschonoského)
- **Malus sect. Eriolobus –** *M. florentina* (jabloň florentinská), *M. trilobata* (jabloň trojlaločná)
- **Malus sect. Gymnomeles** – *M. baccata* (jabloň drobnoplodá), *M. floribunda* (jabloň mnohokvětá), *M. halliana* (jabloň třešňoplodá), *M. hupehensis* (jabloň hupejská), *M. mandshurica* (jabloň mandžuská), *M. sargentii* (jabloň Sargentova), *M. sikkimensis* (jabloň šikimská), *M. spontanea*, *M. toringo* (jabloň Sieboldova)
    - **Hybridy –** *M.* × *arnoldiana* (jabloň Arnoldova), *M.* × *hartwigii* (jabloň Hartwigova), *M.* × *zumi* (jabloň Zumova)
- **Malus sect. Malus** – *M. asiatica*, *M. crescimannoi*, *M. domestica* (jabloň domácí), *M. muliensis*, *M. orientalis* (jabloň východní), *M. prunifolia* (jabloň třešňolistá), *M. pumila* (jabloň nízká), *M. sieversii* (jabloň Sieversova), *M. spectabilis* (jabloň vznešená), *M. sylvestris* (jabloň lesní), *M. zhaojiaoensis*

- **Hybridy –** *M.* × *astracanica* (jabloň astrachánská), *M.* × *eleyi*, *M.* × *magdeburgensis* (jabloň magdeburská), *M.* × *micromalus* (jabloň japonská)

- **Malus sect. Sorbomalus** – *M. fusca* (jabloň hnědá), *M. kansuensis* (jabloň kansunská), *M. komarovii*, *M. maerkangensis*, *M. toringoides* (jabloň laločnatá)
- **Malus sect. Yunnanenses** – M. honanensis (jabloň honanská), M. ombrophila, M. prattii, M. yunnanensis (jabloň junanská)

Malus domestica (jabloň domácí) poprvé popsal a taxonomicky formalizoval německý botanik Moritz Balthasar Borkhausen ve své příručce o lesní botanice (Borkhausen 1803). Mezi známé odrůdy jabloně domácí vyšlechtěné v České republice patří např. Rubín, Admirál, Sonet, Dantes, Karmína, Vltava, Angold, Matčino, Blaník, Lipno, Degas, Čistecké lahůdkové, Aneta, Malinové holovouské, Vysočina, Český ráj, Míšeňské, Atlas, Opál, Bláhovo oranžové, České růžové, Ovčí hubičky, Daria, Panenské české, Hájkova muškátová reneta, Tábor, Diamant, Valašská reneta, Zlaťák a Šampion (https://cs.wikipedia.org/wiki/Seznam_odrůd_jablek).

Druh Malus domestica byl dlouhou dobu považován za hybrid několika druhů jabloní, sám Borkhausen za jeho předchůdce označoval Malus sylvestris (jabloň lesní), Malus dasyphylla (jabloň plstnatá) a Malus praecox (jabloň nízká var. duzén). Podle Hokansona (2001) k jeho vyšlechtění došlo křížením mezi Malus sieversii (jabloň Sieversova), Malus orientalis (jabloň východní), Malus sylvestris, Malus baccata (jabloň drobnoplodá), Malus mandshurica (jabloň bobulovitá) a Malus prunifolia (jabloň třešňolistá). Hypotézu o jediném přímém předchůdci Malus domestica, druhu Malus sieversii, jako první zformuloval sovětský botanik a genetik Nickolai Ivanovich Vavilov (1930), a to na základě podobnosti plodů. Divoká odrůda Malus sieversii pochází z Turkestánu, oblasti rozlehající se od východního pobřeží Kaspického moře až k vrcholům Hindúkuše, Pamíru a Ťan-šanu, odkud se rozšířila do Evropy díky obchodním stezkám vedoucím mezi Evropou a Čínou (viz Obrázek 7). Vavilův předpoklad o příbuzenském vztahu druhů Malus sieversii a Malus domestica výrazně ovlivnil další výzkumy v rámci celého rodu Malus i oblasti centrální Asie jakožto místa diverzifikace odrůd (např. Forsline et al. 1994, Janick et al. 1996, Cornille et al. 2012).




**Obrázek 7**: Malus sieversii pochází z Turkestánu, oblasti v Centrální Asii, kde došlo k diverzifikaci mnoha odrůd, které se rozšířily do ostatních částí světa díky obchodním stezkám mezi Asií a Evropou. (Harris et al. 2002, s. 427)

Zásadní výzkum zabývající se fylogenetickými vztahy mezi jednotlivými druhy a významem centrální Asie pro vývoj kultivarů v rámci rodu Malus vypracovala skupina vědců pod vedením Barrie E. Junipera (1998). O pár let později Juniper spolu se Stephanem Harrisem a Julianem Robinsonem (Harris et al. 2002) tuto studii rozpracovali. Jejich práce spadá do oblasti molekulární fylogenetiky, k analýze využívají molekulární genetické markery, konkrétně chloroplastový marker *matK (maturase K)* a nukleární ribozomální marker *ITS (Internal Transcribed Spacer)*. V případě *matK* byly mezi 21 testovanými druhy Malus identifikovány dva typy duplikace ve vzdálenosti 39-bp (bázových párů) od 3´ konce matK, přičemž polymorfická duplikace o délce 18-bp byla nalezena právě jen u Malus sieversii a Malus domestica. Analýza *ITS* potvrdila silnou příbuzenskou vazbu druhů M. domestica, M. asiatica, M. orientalis, M. prunifolia a M. niedzwetzkyana (pumila) s M. sieversii, přičemž všechny jmenované druhy patří do stejné taxonomické sekce. Získaná data o molekulární variaci podporují předpoklad, že progenitorem domácích jabloní je právě divoká odrůda ze střední Asie, Malus sieversii. Na výzkumy B. E. Junipera navázalo mnoho vědců, pro stejné účely však užili jiné molekulární markery (např. Royo – Itoiz 2004, Gharghani et al. 2009, Gross et al. 2014).

## Lingvistická analýza fylogeneze Malus domestica

Ačkoliv je hypotéza o příbuznosti jabloně domácí s jabloní Sieversovou v současnosti obecně přijímána, sám B. E. Juniper, S. Harris a J. Robinson ve své studii vybízí k dalším analýzám, které by potvrdily či vyvrátily daný předpoklad. Za tímto účelem jsem se rozhodla v praktické části využít metody analýzy popisované v kapitole o lingvistických metodách v bioinformatice. Jako výchozí použiji klasickou bioinformatickou metodu multiple sequence alignment a následně stejný genetický materiál otestuji pomocí Damerau-Levenshteinovy vzdálenosti a Bag-of-Words modelu. Analyzován bude konkrétně nekódující úsek nukleární ribozomální DNA ITS1 5.8S ITS2 a chloroplastový protein matK – gen maturázy K je umístěn uvnitř intronu chloroplastového genu trnK (lysinová tRNA) a kóduje maturázu podílející se na sestřihu RNA transkriptů. Sekvence z nekódujících oblastí chloroplastového genomu se často používají v systematice, protože tyto oblasti mají tendenci se vyvíjet poměrně rychle. Ribosomální RNA je považována za nejlepší cíl pro studium fylogenetického vztahu, protože je univerzální a skládá se z vysoce konzervativních i variabilních domén.

Genetické sekvence, molekulární markery, byly získány z genetické banky NCBI (*National Center for Biotechnology Information*) a Uniprot (*Universal Protein resource*) ve formátu FASTA (*Fast Alignment Search Tool*). Konkrétně byly pro analýzu použity sekvence markerů z 20 druhů: M. asiatica, M. baccata, M. domestica (kultivar Ashmead´s Kernal a Bramley´s Seedling), M. hupehensis, M. micromalus, M. sylvestris, M. toringoides, M. tschonoskii, M. fusca, M. florentina, M. prattii, M. trilobita, M. doumeri, M. ioensis, M. yunnanensis, M. kansuensis, M. doumeri, M. halliana, M. prunifolia. Tyto druhy byly vybrány s cílem otestovat co nejrozmanitější vzorek zástupců různých sekcí rodu Malus a zároveň pro ně bylo možné získat sekvence zvolených markerů. Sekvence (viz přílohy) nebyly pro další analýzy nijak upravovány.

### Multiple sequence alignment

Pro multiple sequence alignment byla jako referenční sekvence zvolena v případě obou markerů sekvence Malus sieversii, aby bylo možné posoudit míru její podobnosti s Malus domestica i ostatními testovanými druhy. Jako typ alignmentu byl zvolen global alignment, sekvence tedy byly přiřazovány po celé délce včetně zanášení mezer (*gaps*) tam, kde byly identifikovány rozdílnosti. Vzhledem k tomu, že oba molekulární markery jsou

zapsány v nukleotidových bázích, v obou případech se penalizace párů a nepárů řídila maticí identity (*identity matrix* IUPAC), kdy shodě byla přidělena penalizace 1 a neshodě penalizace 0. Výsledky alignmentu zaznamenané v tabulkách (viz Tabulka 1, Tabulka 2) prezentují míru podobnosti jednotlivých druhů vzhledem k Malus sieversii a zohledňují také délku jednotlivých sekvencí. Barevně je pak zvýrazněna míra odlišnosti druhů – zelená barva značí odlišnost do 2 %, oranžová barva odlišnost 2–5 % a červená barva odlišnost nad 5 %.

| **Druh** | **Délka** | **shoda** |
|---|---|---|
| M. asiatica | 730 | 730 |
| M. baccata | 730 | 730 |
| M. domestica-Ashmead´s Kernal | 730 | 730 |
| M. domestica-Bramley´s Seedling | 730 | 730 |
| M. prunifolia | 730 | 730 |
| M. hupehensis | 730 | 730 |
| M. micromalus | 730 | 730 |
| M. sylvestris | 730 | 730 |
| M. toringoides | 730 | 730 |
| M. tschonoskii | 730 | 730 |
| M. fusca | 730 | 729 |
| M. florentina | 730 | 728 |
| M. pratii | 730 | 728 |
| M. trilobita | 730 | 728 |
| M. doumeri | 729 | 727 |
| M. ioensis | 730 | 727 |
| M. yunnanensis | 730 | 727 |
| M. kansuensis | 730 | 725 |
| M. coronaria | 728 | 535 |
| M. halliana | 742 | 196 |

**Tabulka 1**: Výsledky globálního alignmentu matK

V případě chloroplastového markeru matK je znatelná menší diverzifikovanost mezi jednotlivými druhy. K naprosté shodě z 20 testovaných sekvencí došlo u 10 druhů a pouze u dvou druhů došlo k výrazné odlišnosti nad 5 % – v případě Malus halliana se jedná o vyšlechtěnou okrasnou odrůdu jabloně a Malus coronaria je odrůda, která byla importována do Severní Ameriky. Jejich výraznější odlišnost tedy můžeme vzhledem k vývojovému charakteru matK přisuzovat především mnohem pozdějšímu vývoji druhů, u Malus halliana také mohla hrát roli kultivace. Pro účel této studie je důležité upozornit na naprostou shodu mezi Malus sieversii a kultivary Malus domestica, což podporuje předpoklad jejich přímé příbuznosti.

| Druh | Délka | shoda |
|---|---|---|
| M. domestica-Ashmead´s Kernal | 591 | 586 |
| M. prunifolia | 591 | 584 |
| M. domestica-Bramley´s Seedling | 591 | 583 |
| M. fusca | 590 | 579 |
| M. hupehensis | 593 | 572 |
| M. toringoides | 592 | 568 |
| M. halliana | 590 | 567 |
| M. coronaria | 593 | 566 |
| M. kansuensis | 593 | 566 |
| M. sylvestris | 603 | 566 |
| M. doumeri | 592 | 564 |
| M. yunnanensis | 593 | 563 |
| M. ioensis | 593 | 561 |
| M. prattii | 593 | 559 |
| M. florentina | 604 | 558 |
| M. trilobita | 593 | 557 |
| M. tschonoskii | 588 | 555 |
| M. asiatica | 613 | 546 |
| M. baccata | 619 | 538 |
| M. micromalus | 614 | 537 |

**Tabulka 2:** Výsledky globálního alignmentu ITS1 5.8S ITS2

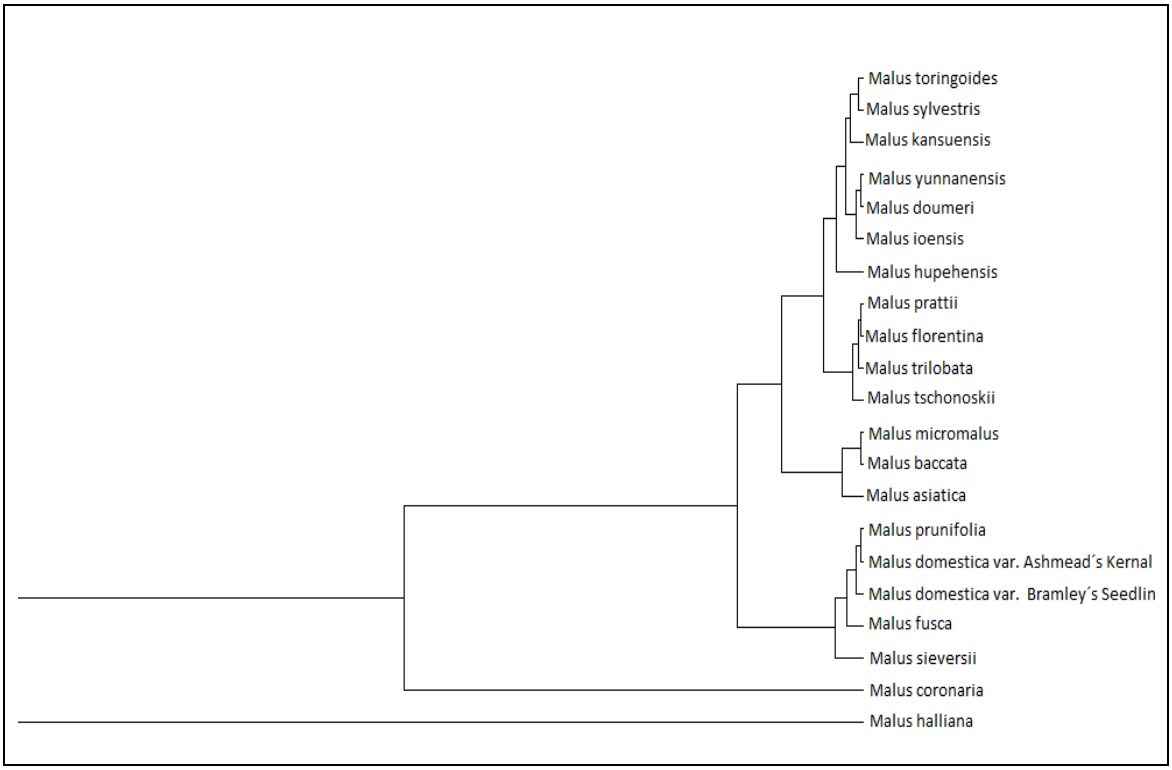
Analýza markeru ITS oproti maturáze K odhalila mnohem větší mezidruhovou variabilitu. Tentokrát nebyla mezi Malus sieversii a Malus domestica identifikována naprostá shoda, ale odrůdy Malus domestica jsou spolu s Malus prunifolia jedinými druhy, jejichž odlišnost od Malus sieversii je do 2 %. Odlišnost 2–5 % byla identifikována u devíti druhů a odlišnost nad 5 % u osmi druhů. Ačkoliv při celkovém zhodnocení výsledků nejsou rozdíly mezi jednotlivými druhy příliš velké, právě variabilita ITS části z celého testovaného úseku ITS1 5.8S ITS2 se ukazuje být rozhodující pro detekování fylogenetických vztahů. Polymorfismus nekódujícího úseku DNA tedy zřetelněji podporuje testovaný předpoklad o evoluční blízkosti domácí a divoké jabloně.

Pokud na základě dat z obou výše uvedených tabulek vytvoříme graf hierarchického shlukování, získáme dendrogram rozdělující druhy do pěti větších větví (viz Obrázek 8):

1. M. toringoides, M. sylvestris, M. kansuensis, M. yunnanensis, M. doumeri, M. ioensis, M. hupehensis, M. prattii, M. florentina, M. trilobita, M. tschonoskii
2. M. micromalus, M. baccata, M. asiatica

3. M. prunifolia, M. domestica var. Bramley´s Seedling, M. domestica var. Ashmead´s Kernal, M. fusca, M. sieversii
4. M. coronaria
5. M. halliana

Na základě této vizualizace dat můžeme pozorovat, že druhy Malus sieversii a Malus domestica jsou ve stejné fylogenetické větvi a vzhledem k velmi krátké délce spojujících kladů (*clades*) můžeme usuzovat jejich blízkou příbuznost. Stejně tak je znatelná blízkost druhů Malus prunifolia a Malus fusca, jejich přímá genetická vazba na Malus sieversii však nebyla prokázána. Důvodem podobnosti těchto dvou druhů s Malus sieversii mohou být obdobné procesy molekulární adaptace, ke kterým došlo vlivem diverzifikace druhů rodu Malus v oblasti Turkestánu, odkud Malus sieversii pochází.




**Obrázek 8**: Dendrogram na základě dat multiple sequence alignment markerů ITS a matK

### Damerau-Levenshteinova vzdálenost

Pro výpočet Damerau-Levenshteinovy vzdálenosti byla pro oba markery jako referenční sekvence použita sekvence Malus sieversii, stejně jako v případě multiple sequence alignmentu. Na základě předpokladu by i tato metoda měla odhalit míru podobnosti Malus domestica a ostatních druhů rodu Malus s referenční Malus sieversii. V tomto případě však metoda nesleduje, jak jsou si sekvence podobné, ale jak jsou si vzdálené – nehledá tedy úseky sekvencí, které jsou totožné, ale zaměřuje se na identifikaci změn a jejich povahu. Jak bylo popsáno výše, změnami/transformacemi jsou myšleny druhy bodových mutací, kterými se sekvence od sebe navzájem liší: *delece* (chybějící část), *inverze/substituce* (záměna částí), *inzerce* (začlenění části) nebo *transpozice* (prohození dvou sousedících částí). Pro každou z těchto transformací byla zvolena penalizace 1, pro shodu bází penalizace 0.

Výsledky analýzy Damera-Levenshteinovy vzdálenosti jsou zaznamenány v tabulce pro každý marker zvlášť (viz Tabulka 3, Tabulka 4). Zatímco u alignmentu vyšší hodnota vypovídala o větší podobnosti druhové sekvence s Malus sieversii, u této metody vyšší hodnota znamená, že došlo k více transformacím, tudíž je druhová sekvence od Malus sieversii vzdálenější. V tabulce je vždy opět barevně zvýrazněna míra odlišnosti druhů – zelená barva značí odlišnost do 2 %, oranžová barva odlišnost 2–5 % a červená barva odlišnost nad 5 %.

U maturázy K je opět znatelné velké zastoupení druhů s odlišností (vzdáleností) do 2 %, z dvaceti vzorků se jedná o 14 druhů. Pět druhů pak vykazuje odlišnost 2–5 % a pouze jeden druh se vyznačuje větší vzdáleností od Malus sieversii – jedná se o okrasnou odrůdu Malus halliana, která se i v multiple sequence alignment analýze výrazně odchýlila od ostatních druhů. Můžeme si všimnout, že variety druhu Malus domestica jsou v tabulce na druhém a třetím místě. Konkrétně u variety Bramley´s Seedling došlo pouze ke dvěma transformacím a u variety Ashmead´s Kernal ke čtyřem transformacím. Získané výsledky vypovídají o téměř absolutní shodě mezi sekvencemi maturázy K u druhu Malus sieversii a variet druhu Malus domestica. Tím podporují testovaný předpoklad o přímém fylogenetickém vztahu mezi těmito druhy.

| Sekvence | délka | vzdálenost |
|---|---|---|
| M. sylvestris | 745 | 0 |
| M. domestica-Bramley´s Seedling | 743 | 2 |
| M. domestica-Ashmead´s Kernal | 741 | 4 |
| M. tschonoskii | 741 | 4 |
| M. micromalus | 737 | 8 |
| M. hupehensis | 736 | 9 |
| M. trilobita | 737 | 10 |
| M. ioensis | 737 | 11 |
| M. yunnanensis | 737 | 11 |
| M. asiatica | 733 | 12 |
| M. fusca | 733 | 13 |
| M. toringoides | 732 | 13 |
| M. florentina | 733 | 14 |
| M. prunifolia | 731 | 14 |
| M. coronaria | 733 | 15 |
| M. prattii | 733 | 15 |
| M. doumeri | 733 | 16 |
| M. kansuensis | 733 | 18 |
| M. baccata | 736 | 19 |
| M. halliana | 745 | 364 |

**Tabulka 3**: Damerau-Levenshteinova vzdálenost maturázy K

| Sekvence | délka | vzdálenost |
|---|---|---|
| M. domestica-Ashmead´s Kernal | 594 | 7 |
| M. prunifolia | 593 | 10 |
| M. domestica-Bramley´s Seedling | 594 | 11 |
| M. fusca | 593 | 14 |
| M. hupehensis | 596 | 24 |
| M. halliana | 594 | 27 |
| M. toringoides | 594 | 27 |
| M. doumeri | 597 | 29 |
| M. coronaria | 597 | 30 |
| M. kansuensis | 596 | 32 |
| M. ioensis | 597 | 33 |
| M. yunnanensis | 597 | 35 |
| M. sylvestris | 604 | 37 |
| M. tschonoskii | 594 | 37 |
| M. prattii | 595 | 38 |
| M. trilobita | 596 | 40 |
| M. asiatica | 618 | 52 |
| M. florentina | 609 | 57 |
| M. baccata | 626 | 58 |
| M. micromalus | 618 | 63 |

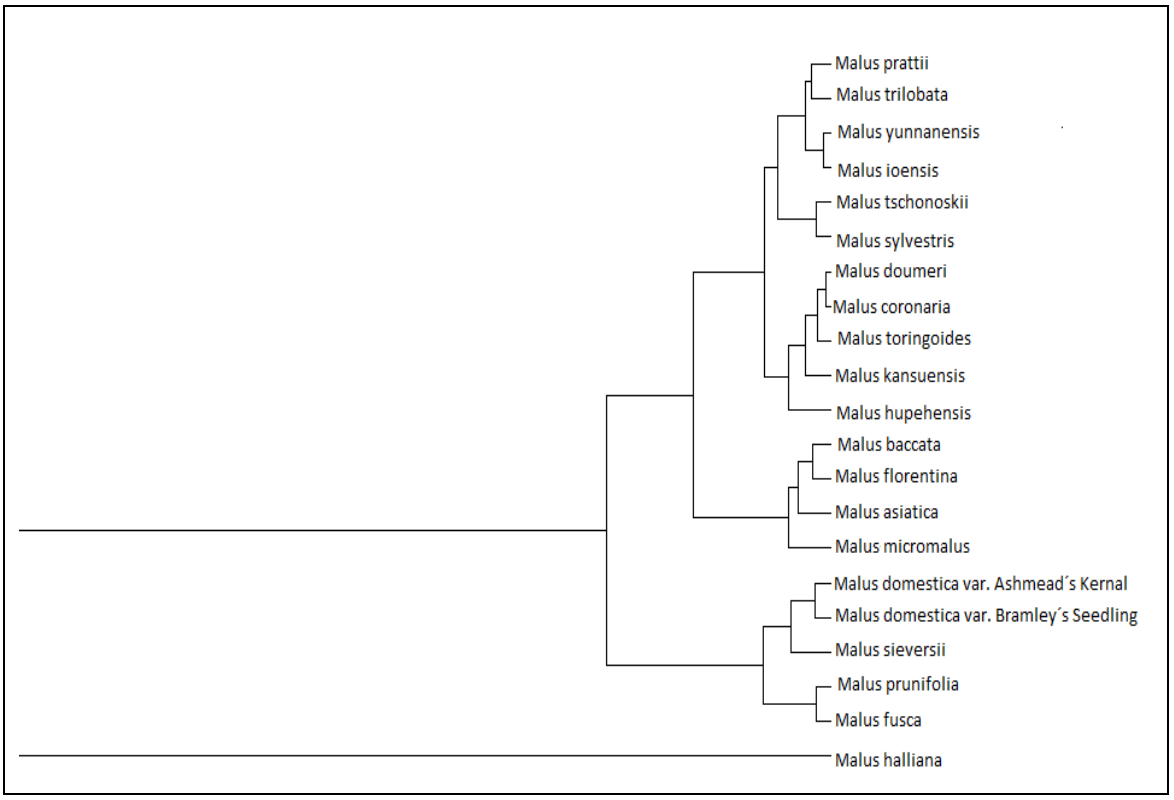
**Tabulka 4**: Damerau-Levenshteinova vzdálenost ITS1 5.8S ITS2

Analýza nukleárního ribozomálního markeru ITS díky své variabilitě znovu poskytla rozmanitější výsledky. Dokonce lze tvrdit, že poměr druhů s odlišností do 2 % a druhů s odlišností nad 5 % je oproti výsledkům analýzy markeru matK opačný. Zatímco do skupiny nejodlišnějších druhů patří 12 druhů, mezi nejpodobnější patří pouze 3. Pozoruhodné je, že mezi tyto nejpodobnější druhy patří variety Malus domestica a Malus prunifolia, která je v tabulce na druhém místě mezi těmito varietami. Stejný výsledek, co se týče umístění v tabulce, byl získán při analýze markeru ITS pomocí multiple alignmentu. I zde tedy lze hovořit o důkazu podporujícím hypotézu o příbuznosti Malus sieversii a Malus domestica.

Pokud porovnáme tabulky výsledků z obou analýz, multiple sequence alignmentu a Damerau-Levenshteinovy distance, pro každý marker zvlášť, můžeme posoudit citlivost použitého markeru na zvolenou metodu vzhledem k výzkumnému cíli – zjištění míry podobnosti, a tedy příbuznosti zkoumaných druhů. V případě markeru matK jsou u obou analýz variety Malus domestica vyhodnoceny jako zcela nebo téměř naprosto shodné s Malus sieversii. Ostatní druhy mají v každé tabulce rozdílná umístění. Zajímavé však je, že v případě analýzy markeru ITS obě metody poskytují výsledky s výrazně podobným pořadím, navíc u pozic variet Malus domestica jsou naprosto shodné. Každá z těchto metod zaujímá odlišný přístup k zhodnocení míry podobnosti/odlišnosti mezi dvěma sekvencemi, což vede ke dvěma zjištěním: Zaprvé, rozdílnost výsledků při analýzách markeru matK, co se týče rozmístění druhů v rámci spektra odlišnosti, poukazuje na větší citlivost tohoto markeru vůči použité metodě. A zadruhé, analýza obou markerů oběma metodami vždy poskytla výsledky podporující předpoklad příbuznosti divoké a domácí jabloně.

Na základě dat z obou tabulek opět vytvoříme graf hierarchického shlukování, který lépe vizualizuje fylogenetické vztahy druhů. V tomto dendrogramu (viz Obrázek 9) můžeme identifikovat čtyři větší fylogenetické větve:

1. M. prattii, M. trilobita, M. yunnanensis, M. ioensis, M. tschonoskii, M. sylvestris, M. doumeri, M. coronaria, M. toringoides, M. kansuensis, M. hupehensis
2. M. baccata, M. florentina, M. asiatica, M. micromalus
3. M. domestica (kultivary Ashmead´s Kernal a Bramley´s Seedling), M. sieversii, M. prunifolia, M. fusca
4. M. halliana




**Obrázek 9**: Dendrogram na základě dat Damerau-Levenshteinovy vzdálenosti markerů matK a ITS

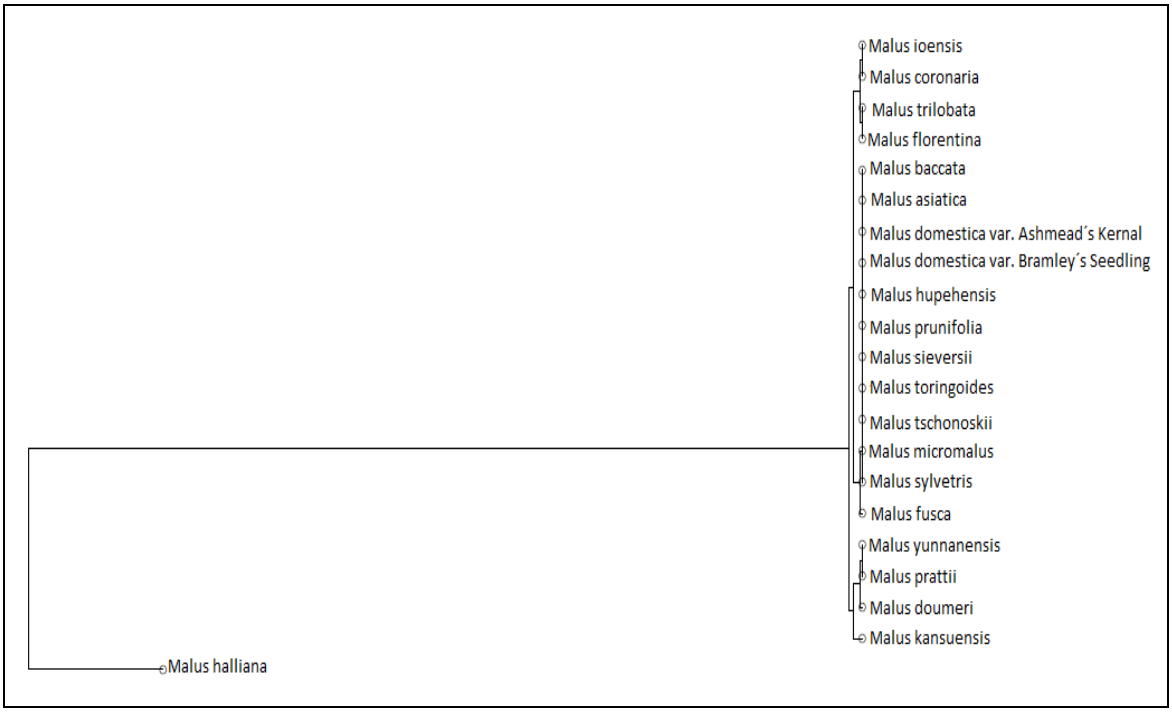
Jak je z grafu patrné, Malus sieversii a odrůdy Malus domestica jsou spolu s Malus prunifolia a Malus fusca v jedné z větších fylogenetických větví. Stejného shluku těchto druhů do společné větve bylo dosaženo také v dendrogramu vytvořeného na základě dat z multiple sequence alignmentu. Když vezmeme v úvahu pouze nejnižší úroveň větvení, tak Malus sieversii a Malus domestica byly začleněny do jednoho shluku (klastru), což jen potvrzuje jejich příbuznost uvedenou již v tabulkách. Vlivem citlivosti markeru matK na použitou metodu byly detekovány rozdíly v pořadí druhů v tabulkách, tyto odlišnosti zohledňuje také vizualizace metodou hierarchického shlukování. Při porovnání obou dendrogramů nalezneme jisté podobnosti ve shlukování jednotlivých druhů, naprostou shodou zůstává oddělení M. halliana od ostatních druhů do samostatné větve a shluk druhů M. sieversii, M. domestica, M. prunifolia a M. fusca do jedné větve.

## Bag-of-words model

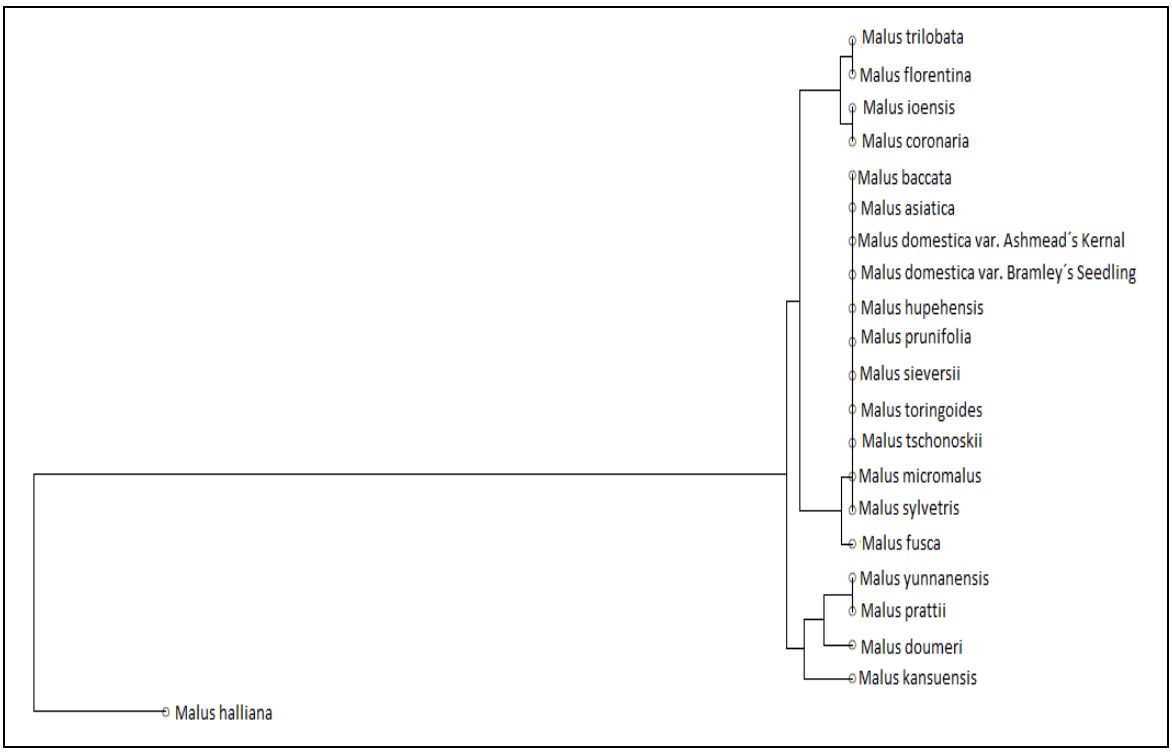
Jak jsem vysvětlila výše, tzv. BoW model spočívá v reprezentaci textu jeho slovy, která jsou v případě genetických textů určena n-gramovou analýzou. Zohledněna je frekvence slov a také multikódový charakter genetických sekvencí (předpokládáme-li, že genetické sekvence kódují různé funkce a v jejich sekvenci se jejich zápis překrývá, pak je n-gramová analýza vhodná k registraci těchto překryvů, protože okno výběru n-gramu postupuje po sekvenci vždy po jedné jednotce – aminokyselině/bázi). Pro účely této analýzy byl využit program QUITA (*Quantitative Index Text Analyzer*), kde byly sekvence markerů tokenizovány (tj. rozčleněny na jednotlivé tokeny, v našem případě nukleotidové báze) a dále rozděleny na požadované n-gramy, tedy úseky stejné délky reprezentující slova. Konkrétně byly sledovány změny, ke kterým dojde, pokud genetický slovník markerů bude reprezentován slovy v podobě 3-gramů, 5-gramů a 10-gramů – počet gramů ovlivňuje podobu slov, pravděpodobnost jejich výskytu v genetickém textu a jejich frekvenci. Tímto postupem byly jednotlivé markery reprezentovány přítomností či nepřítomností definovaných slov a jejich četností. Díky tomu mohly být markery mezi sebou porovnávány na základě podobnosti svých genetických slovníků. Jelikož použité sekvence nemají všechny stejnou délku, byla pro měření vzdálenosti mezi druhy použita cosinova distance. Pro vizualizaci podobnosti sekvencí bylo zvoleno hierarchické shlukování a MDS (*Multidimensional Scaling*).

Předchozí analýzy ukázaly, že maturáza K vykazuje mezi druhy jen nepatrné rozdíly, z čehož lze usuzovat na výraznou podobnost jejich slovníků, což Bag-of-Words model potvrdil. Jelikož při změně parametru n-gramu nedocházelo k výrazným odlišnostem, jsou zde prezentovány grafy dendrogramu pouze pro 3-gramovou a 10-gramovou reprezentaci slov (viz Obrázek 10 a Obrázek 11). Jak je patrné, shluky druhů do hlavních větví jsou u obou grafů stejné, pouze u 10-gramů se zvětšila délka kladů (dílčích vývojových větví). Můžeme zde identifikovat 4 větší větve:

1. M. ioensis, M. coronaria, M. trilobata, M. florentina
2. M. baccata, M. asiatica, M. domestica (varieta Ashmead´s Kernal, Bramley´s Seedling), M. hupehensis, M. prunifolia, M. sieversii, M. toringoides, M. tschonoskii, Malus micromalus, Malus sylvestris, M. fusca
3. M. yunnanensis, M. prattii, M. doumeri, M. kansuensis
4. M. halliana




**Obrázek 10**: Dendrogram Bag-of-Words modelu matK markeru, 3-gram



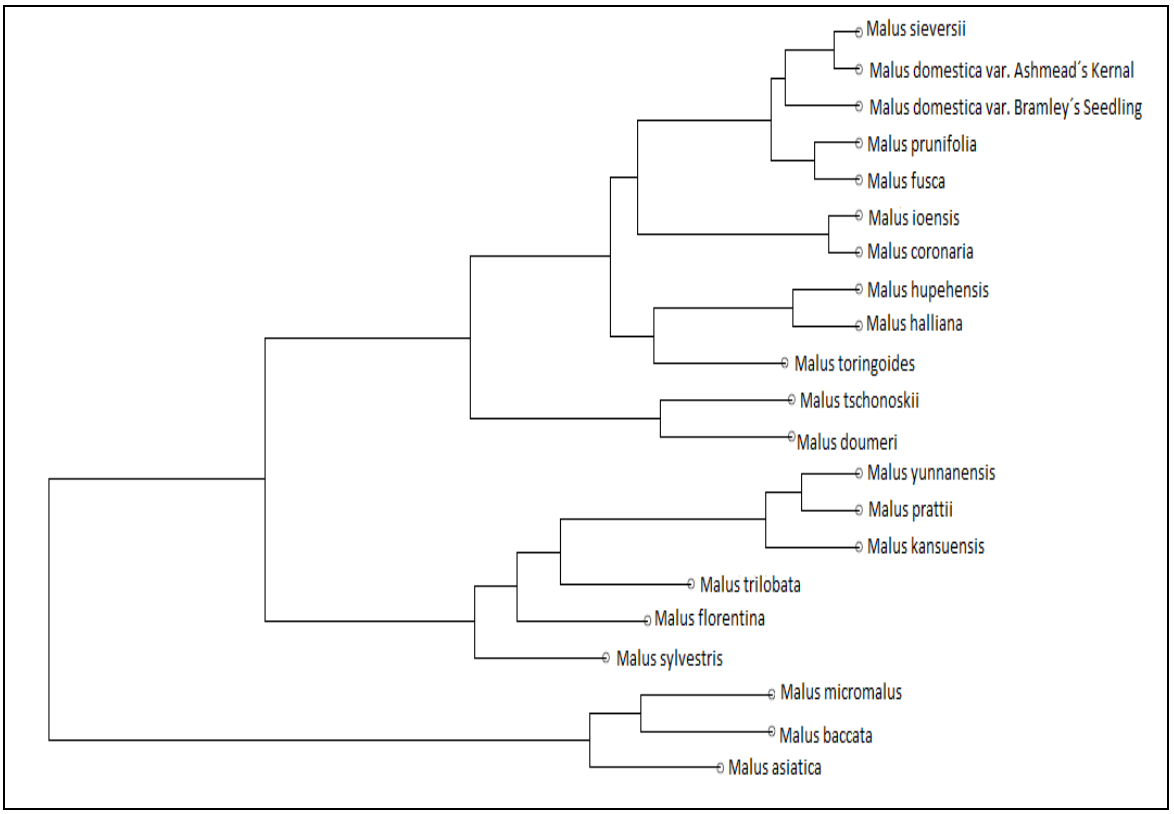

**Obrázek 11**: Dendrogram Bag-of-Words modelu matK markeru, 10-gram

Clustrování na základě Bag-of-Words modelu potvrdilo podobnost matK markeru u jednotlivých druhů, kterou můžeme pozorovat u multiple sequence alignmentu a Damerau-Levenstheinovy vzdálenosti (viz výsledky v tabulkách). Ve třetí velké větvi můžeme vzhledem k pozici kladů v jedné rovině vidět naprostou shodu mezi Malus sieversii a odrůdami Malus domestica, ale taky dalšími druhy, jak výše uvádí Tabulka 1. Patrné je také výrazné oddělení Malus halliana do samostatné větve, což poukazuje na její velkou odlišnost od ostatních druhů, taktéž zaznamenanou v tabulkách multiple sequence alignmentu a Damerau-Levenstheinovy vzdálenosti. Kvůli malým rozdílům ve slovnících jednotlivých druhů zde není prezentována vizualizace pomocí MDS, jelikož se druhy navzájem překrývají a není tedy příliš přehledná.

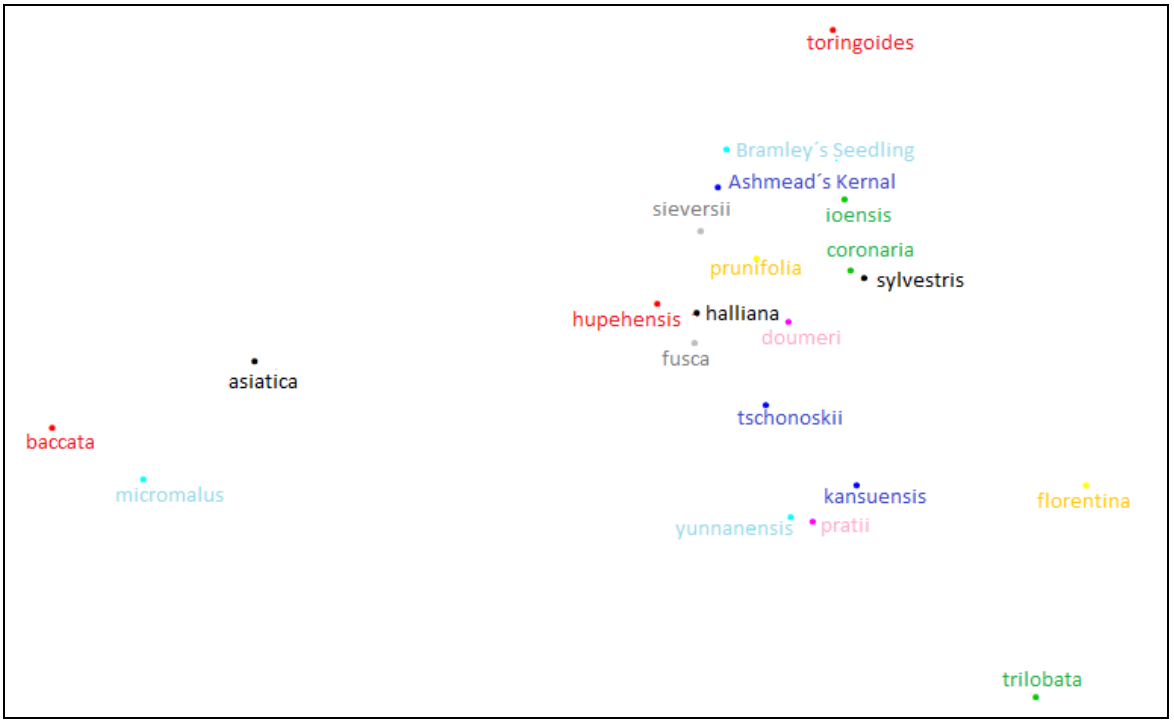
Dříve jsme v analýze identifikovali mnohem větší variabilitu u ribozomálního markeru ITS než u chloroplastového markeru matK. Tento jeho rys lze pozorovat také v Bag-of-Words analýze, pro názornost jsou uvedeny grafy hierarchického shlukování i MDS pro všechny tři testované n-gramy. V prvním dendrogramu můžeme identifikovat devět větších větví:

1. M. hupehensis, M. halliana, M. toringoides
2. M. sieversii, M. domestica (variety Ashmead´s Kernal, Bramley´s Seedling), M. fusca, M. prunifolia
3. M. trilobita
4. M. florentina
5. M. sylvestris
6. M. ioensis, M. coronaria
7. M. tschonoskii, M. doumeri
8. M. yunnanensis, M. prattii, M. kansuensis
9. M. baccata, M. asiatica, M. micromalus

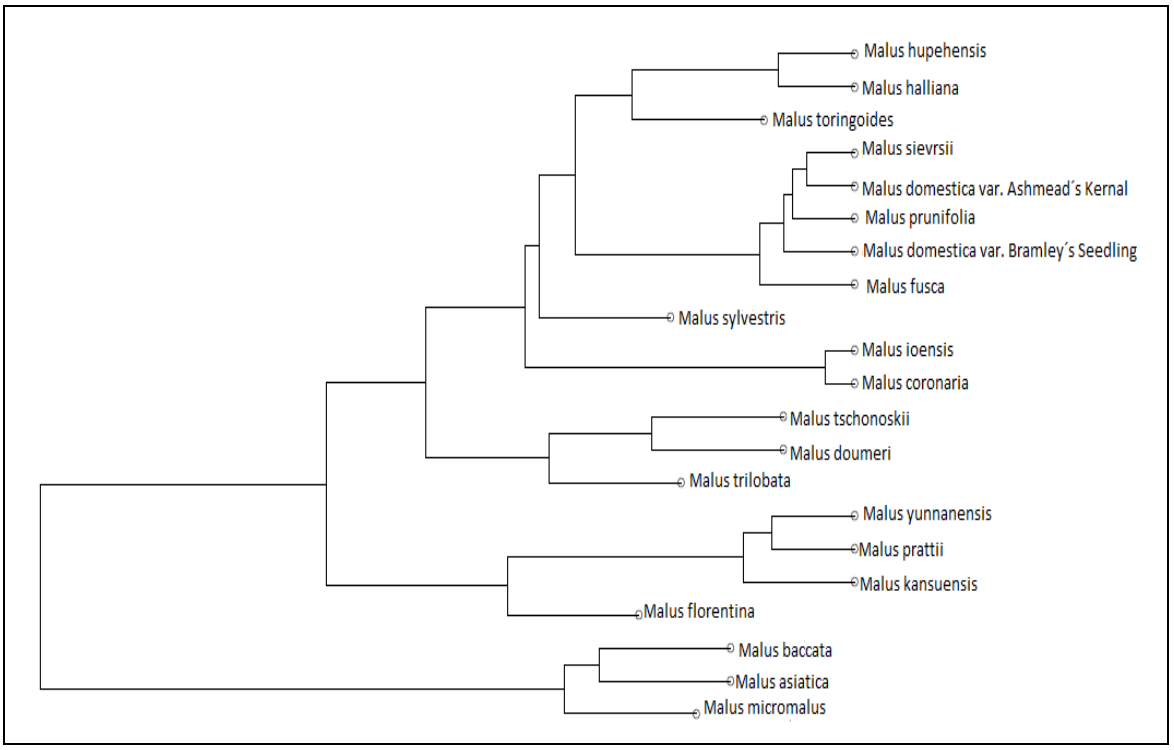
Toto clustrování tedy platí pro 3-gramy, u 5-gramů zaniknou dvě samostatné větve a připojí se k jiným – M. florentina se připojí k druhům ve větvi označené číslem 8 a M. trilobita k druhům ve větvi označené číslem 7. V grafu 10-gramů se pak M. florentina odpojí od 8. větve a přičlení se k 7. větvi. Kromě změn týkajících se těchto dvou druhů je shlukování druhů do větví i při změnách n-gramů zachováno, mění se však jejich vertikální pořádí v grafu. Další změnou, kterou můžeme sledovat, je postupné prodlužování kladů na nejnižší úrovni větvení (viz Obrázek 12, Obrázek 14 a Obrázek 16).




**Obrázek 12**: Dendrogram Bag-of-Words modelu ITS markeru, 3-gram

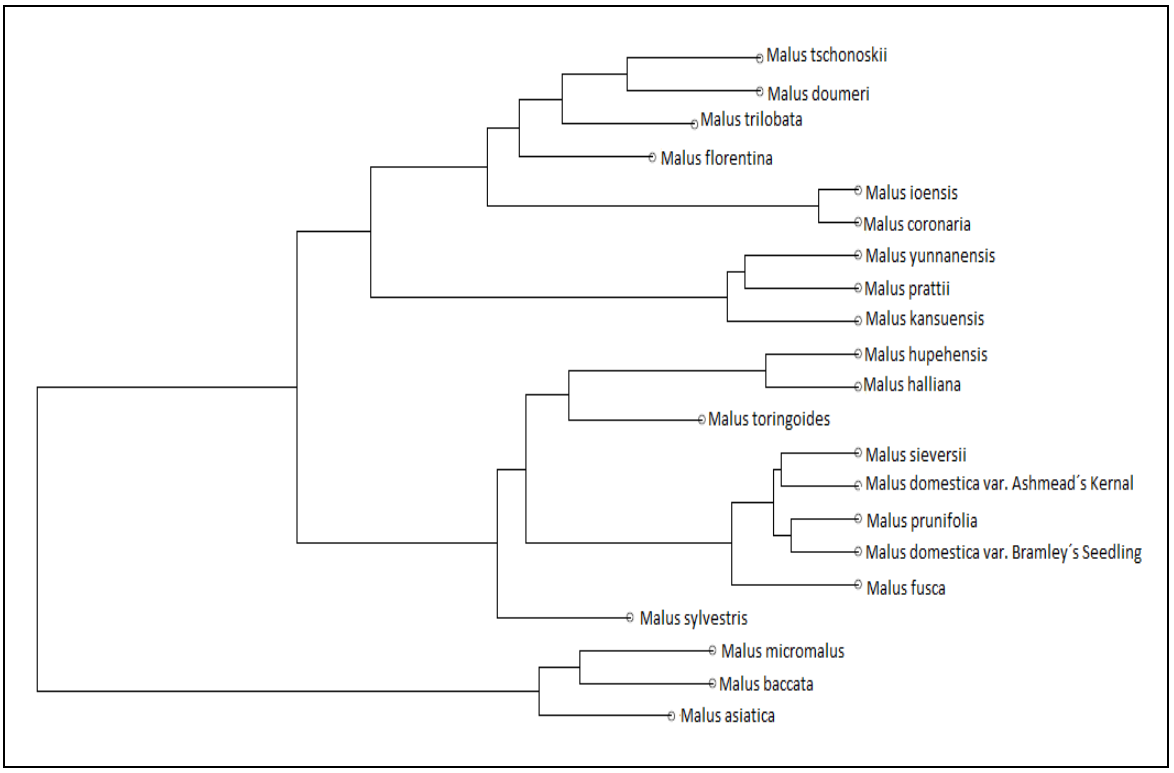



**Obrázek 13**: MDS Bag-of-Words modelu ITS markeru, 3-gram

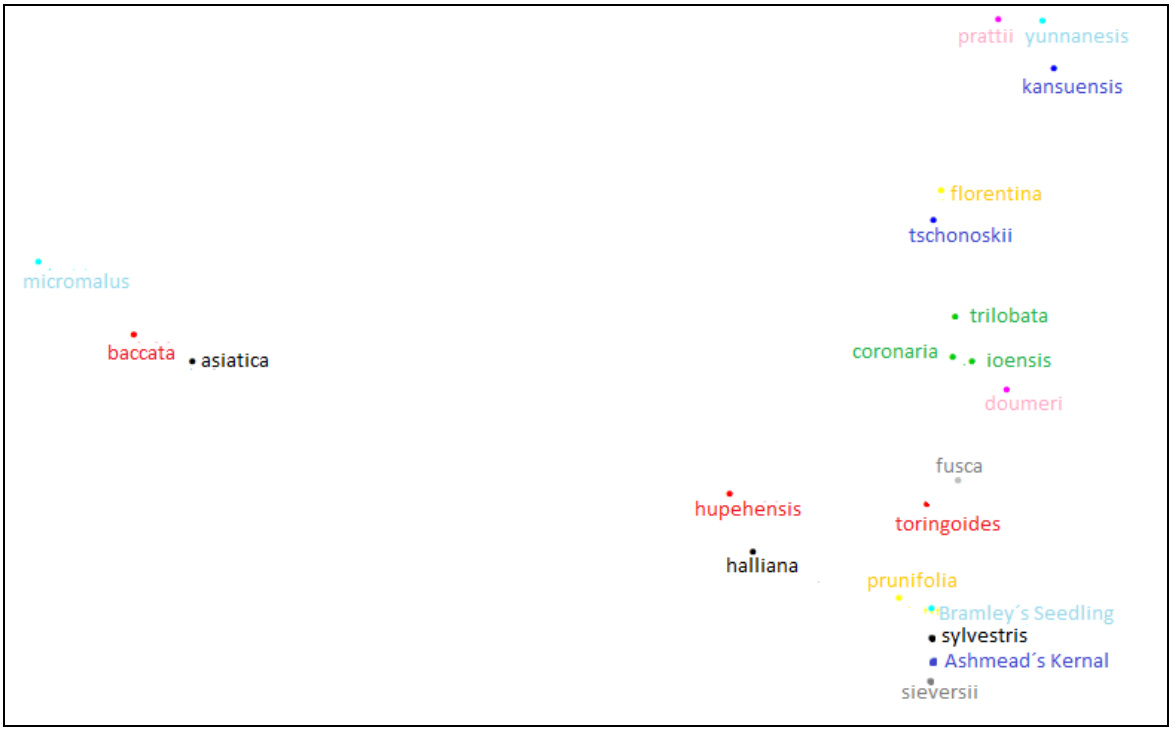



**Obrázek 14**: Dendrogram Bag-of-Words modelu ITS markeru, 5-gram


**Obrázek 15**: MDS graf Bag-of-Words modelu ITS markeru, 5-gram




**Obrázek 16**: Dendrogram Bag-of-Words modelu ITS markeru, 10-gram




**Obrázek 17**: MDS graf Bag-of-Words modelu ITS markeru, 10-gram

Pozorujeme-li probíhající změny zaznamenané mezi jednotlivými dendrogramy, můžeme usuoudit, že délka genetického slova (n-gramu) ovlivňuje clustrování, které vypovídá o podobnosti jednotlivých druhů. Při změnách délky slova (zvětšování n-gramů) došlo k přeskupení některých druhů a postupnému prodlužování kladů – z toho tedy vyplývá, že při zkoumání kratších slov nalezneme mezi sekvencemi více podobností než u delších slov. Pokud se vrátíme zpět k metafoře mezi přirozeným a genetickým jazykem, mohli bychom tento jev vysvětlit tak, že jednotlivé slabiky nebo kořeny slov budou společné mnoha textům, ale když ze slabik složíme slova nebo ke kořenům přidáme různé afixy, podobnost mezi texty se začne zmenšovat až vytrácet.

Vedle změn v hierarchickém shlukování můžeme také sledovat, jak se mění shluky v MDS grafech (Obrázek 13, Obrázek 15 a Obrázek 17). Tato vizualizace nám umožní vidět vztahy mezi druhy prostorově, oproti stromovému větvení dendrogramu. Při analýze využívající 3-gramy je patrné, že všechny druhy jsou v jednom velkém shluku, kromě tří druhů tvořících větev označenou číslem 9 (M. asiatica, M. baccata, M micromalus), která je i v dendrogramu zcela oddělena od ostatních větví. U 5-gramů se druhy v tomto shluku ještě více semknou a můžeme sledovat, jak se postupně i ostatní druhy oddělují s původně jednoho velkého shluku. V případě 10-gramů je pak jednoznačně viditelné rozdělení druhů do shluků, které odpovídají jednotlivým větvím dendrogramu. Vizualizace pomocí MDS nám tedy ještě lépe zobrazuje, jak délka slova ovlivňuje diverzifikaci druhů.

Vzhledem k účelu této práce se především zaměříme, v jakém vztahu je Malus sieversii a Malus domestica. Při rozdělení druhů do větších větví byly oba druhy zařazeny do jedné větve, označené číslem 2, společně s Malus prunifolia a Malus fusca. Můžeme si povšimnout, že tyto druhy zůstávají ve stejném clusteru ve všech dendrogramech, a to ve všech provedených analýzách (BoW, multiple sequence alignment, Damerau-Levenshteinova vzdálenost). V BoW modelu se změnou n-gramů dochází ke zmíněnému postupnému prodlužování kladů a Malus prunifolia změní pozici přesunutím mezi variety Malus domestica. Ve všech grafech hierarchického shlukování zůstávají variety M. domestica v těsné blízkosti M. sieversii. Grafy MDS ukazují posuzované druhy vždy ve vzájemné blízkosti a při větších n-gramech se oddělují od ostatních druhů – v MDS grafu 10-gramů je můžeme vidět v samostatném shluku společně s M. prunifolia, se kterou sdílejí větev, a s M. sylvestris, která je v dendrogramu v sousední větvi. Je tedy patrné, že Bag-of-

Words analýza i při změně testovaného markeru (matk, ITS), parametrů (n-gramů) a vizualizace mezidruhové distance (hierarchické shlukování, MDS) potvrzuje předpoklad o přímém příbuzenském vztahu divoké Sieversovy a domácí jabloně.

## Závěr

Cílem této kapitoly bylo v návaznosti na výzkumy B. Junipera, S. Harrise a J. Robinsona otestovat hypotézu o fylogenetické příbuznosti druhů Malus sieversii a Malus domestica. K tomu byla využita standartní bioinformatická metoda multiple sequence alignment a lingvistické metody, jejichž využití v bioinformatice není standardizované, Damerau-Levenshteinova vzdálenost a Bag-of-Words model. Testovány byly molekulární markery, konkrétně chloroplastový marker maturáza K a jaderný ribozomální marker ITS1 5.8S ITS2. Všechny metody analýzy na vybraném genetickém materiálu přinesly výsledky podporující předpoklad, že divoká jabloň Sieversova z oblasti Tyrkystánu v centrální Asii je přímým předchůdcem jabloně domácí. České přísloví „jablko nepadá daleko od stromu“ tedy v tomto případě nedošlo svého naplnění. Zároveň je tím prokázána využitelnost lingvistických metod, které pracují s genetickým materiálem jako s textem, pro výzkumy spadající do oblasti molekulární fylogenetiky.

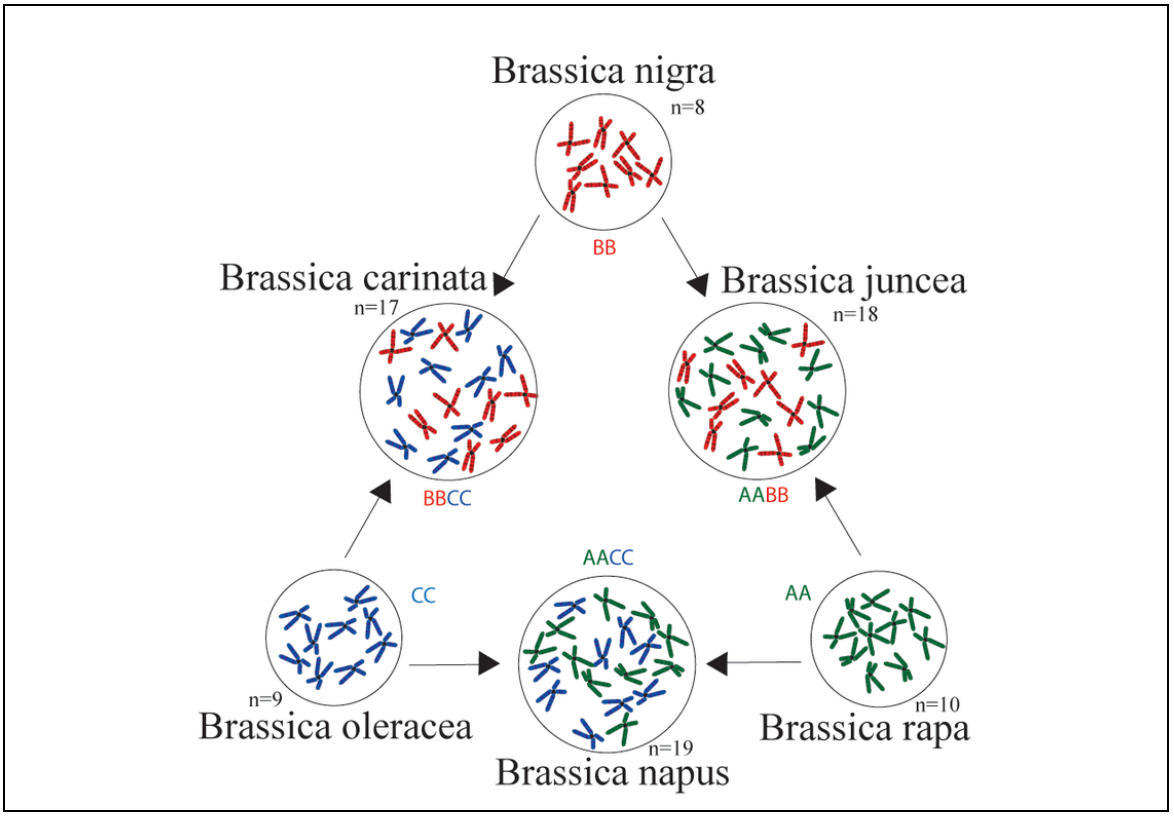
# Fylogenetické vztahy v rámci rodu Brassica

*„Cauliflower is nothing but cabbage with a college education."* *Mark Twain*

Rod Brassica (brukev) na základě botanické klasifikace patří do řádu brukvotvaré (Brassicales), čeledi brukvovité (Brassicaceae) a tribu Brassiceae. Jeho zástupci jsou známí především pro své rozsáhlé hospodářské využití, přesto do současnosti neexistuje jednotná systematika popisující počet a zařazení jednotlivých taxonů – např. server BrassiBase, systém pro taxonomii a mapování fylogenetických vztahů čeledi Brassicaceae (https://brassibase.cos.uni-heidelberg.de), uvádí 44 druhů rodu Brassica včetně jejich variet. Mezi nejvýznamnější kulturní plodiny tohoto rodu patří B. nigra (brukev černá/černá hořčice), B. juncea (brukev sítinovitá/hořčice), B. carinata (brukev kýlnatá/hořčice habešská), B. rapa (brukev řepák), B. oleracea (brukev zelná), B. napus (brukev řepka olejka), B. oleracea var. botrytis (květák), B. oleracea var. italica (brokolice), B. oleracea var. capitata (hlávkové zelí), B. oleracea var. alboglabra (čínská brokolice), B. oleracea var. acephala (kapusta kadeřavá), B. oleracea var. gemmifera (růžičková kapusta), B. oleracea var. gongylodes (kedluben) B. napus var. napobrassica (brukev řepka tuřín), B. rapa var. chinensis (čínské zelí), B. rapa var. pekinensis (pekingské zelí), B. rapa var. oleifera (brukev řepák olejný), B. campestris rapifera (neboli B. rapa var. rapa, brukev řepák vodnice).

Významnou prací mapující příbuzenské vztahy v rámci rodu Brassica bylo popsaní tzv. U modelu (*Triangle of U*) Woo Jang-Choonem, neboli Nagaharu U (1935). Tento model (viz Obrázek 18) popisuje vztahy a vývoj tří druhů, jejichž spojením vznikly nové druhy zeleniny a olejnin; ukazuje, že tři druhy rodu Brassica byly odvozeny ze tří rodových genomů označených písmeny *AA*, *BB* a *CC* (písmeno *n* označuje počet chromozomů). Každý z těchto diploidních genomů tvoří samostatný druh, ale kvůli jejich blízké příbuznosti je bylo možné zkřížit, a tak vytvořit tři nové tetraploidní druhy. Protože jsou odvozeny z genomů dvou různých druhů, jsou tyto hybridní rostliny označovány jako allotetraploidní, tj. obsahují čtyři genomy odvozené ze dvou různých rodových druhů. Konkrétněji jsou amfidiploidní, tj. obsahují jeden diploidní genom z každého ze dvou různých druhů Brassica (Lysák 2007). Dodnes představuje U model základní orientační schéma

křížení mezi jmenovanými druhy a byl potvrzen studiem DNA a proteinů (např. Schmidt – Bancroft 2011).




**Obrázek 18**: Diagram U modelu zobrazující genetické vztahy šesti druhů rodu Brassica (wikipedia.org)

**AA – 2n=2x=20** – Brassica rapa
**BB – 2n=2x=16** – Brassica nigra
**CC – 2n=2x=18** – Brassica oleracea
**AABB – 2n=4x=36** – Brassica juncea
**BBCC – 2n=4x=34** – Brassica carinata
**AACC – 2n=4x=38** – Brassica napus

Průlomovým objevem bylo historicky první zmapování genomu rostliny, a to druhu Arabidopsis thaliana (huseníček rolní) v roce 2000 zásluhou projektu zvaného *Arabidopsis Genome Initiative* (www.arabidopsis.org). Jedná se o rostlinu z čeledi Brassicaeae s krátkou generační dobou a poměrně malým jaderným genomem (157 milionů párů bází), která slouží jako modelová referenční rostlina ve fylogenetických studiích. Pro zkoumání fylogenetických vztahů v rámci rodu Brassica (ale také celé čeledi Brassicaeae) je běžně využívána (např. Parkin et al. 2005, Town et al. 2006, Yamamoto – Nishio 2014)

Taxonomie čeledi Brassicaeae byla systematicky popsána a revidována mnoha autory na základě metod molekulární fylogenetiky pro jednotlivé triby a rody, přehled o těchto výzkumech podává Al-Shehbaz (2006). V českém prostředí se studiu rodu Brassica a šířeji

celé čeledi Brassicaeae věnuje výzkumný tým pod vedením Martina Lysáka, který pracuje ve vědeckém centru CEITEC (*Central European Institute of Technology)*. Na organizaci centra se podílí šest významných brněnských univerzit a výzkumných institucí; mezi oblasti zájmu tohoto centra patří kromě genomiky a proteomiky rostlinných systémů také nanotechnologie, mikrotechnologie, pokročilé materiály, strukturní biologie, molekulární medicína a výzkum mozku (www.ceitec.cz). Lysák se se svými kolegy zaměřuje především na analýzu struktury chromozomů, změny v centromerech, karyotypové variace, vývoj repetitivních sekvencí a rekonstrukci rodových genomů čeledi Brassicaeae (např. Lysák 2009, Franzke et al. 2011, Lysák – Koch 2011, Kiefer et al. 2014, Cheng et al. 2015).

Pro genomickou a genetickou analýzu rodu Brassica bylo v minulosti využito mnoho bioinformatických a biochemických nástrojů a informačních zdrojů, další jsou v současnosti zkoumány a rozvíjeny. Jednou z největších výzev pro budoucí výzkumy je sekvenování genomů variet a divokých druhů pro identifikaci fenotypových variací. Tyto informace by měly posloužit rozvoji pěstování hospodářsky využívaných plodin. Další nezmapovanou oblastí je způsob, jakým byla přenesena dědičná informace u hybridů od jejich progenitorů. Na cestě k dosažení pokroku v těchto oblastech vyvstává potřeba nalezení nových přístupů k analýze genetických dat.

## Lingvistická analýza fylogenetických vztahů v rámci U-modelu

Výše popsaný U-model popisující vztahy tří druhů rodu Brassica a hybridů, které vznikly jejich zkřížením, je od svého vzniku v roce 1935 uznáván jako platné výchozí schéma pro výzkumy dalších fylogenetických vztahů v rámci tohoto rodu a také jemu taxonomicky nadřazeného tribu Brassiceae. Jak bylo dříve uvedeno, právě povaha přenosu dědičné informace z progenitorů na jejich hybrid není do současnosti zcela objasněna. V návaznosti na výzkumy týmu Martina Lysáka v centru CEITEC i výzkumy provedené bioinformatiky v zahraničí bych chtěla v této části provést lingvistickou analýzu fylogenetických vztahů druhů tvořících U-model a rozšířit jej o variety těchto druhů. Cílem tohoto přístupu je zjistit, zda jsou lingvistické metody vhodným nástrojem pro mapování také složitějších vývojových vztahů, a zároveň rozšířit výklad o vztahy původních plodin ke svým kulturně využívaným varietám. Pro analýzu bude opět využita metoda Damerau-Levenshteinovy vzdálenosti a Bag-of-Words model. Ačkoliv jsou genetické markery jaderné ribozomální DNA a chloroplastových proteinů běžně ve fylogenetických analýzách rostlin využívány, kombinace markeru ITS a matK v případě analýzy druhů rodu Brassica zatím využita nebyla. Proto využiji pro nové metody v rámci molekulární fylogenentiky právě markery ITS1 5.8S ITS2 a matK, které u jiných druhů kombinovány byly a osvědčily se v analýze rodu Malus.

Genetické sekvence, molekulární markery, byly získány z genetické banky NCBI (*National Center for Biotechnology Information*) a Uniprot (*Universal Protein resource*) ve formátu FASTA (*Fast Alignment Search Tool*). Pro porovnání výsledků z analýz jednotlivých markerů je nutné, aby byly sekvence obou markerů zapsány v bázích. U ITS markeru bylo možné získat sekvence variet zkoumaných druhů, ale u maturázy K to možné nebylo; analýzy se tedy ve vzorku liší, ale je možné porovnat vztahy druhů tvořících základní kostru U-modelu, které jsou v analyzovaném vzorku vždy přítomny. V analýzách je tedy pracováno s druhy B. oleracea, B. rapa, B. nigra, B. juncea, B. carinata, B. napus, a jejich varietami B. oleracea var. acephala, B. oleracea var. alboglabra, B. oleracea var. capitata, B. oleracea var. botrytis, B. rapa var. chinensis, B. rapa var. pekinensis, B. rapa var. oleifera a B. campestris rapifera (B. rapa var. rapa). Sekvence (viz přílohy) nebyly pro další analýzy nijak upravovány.

## Damerau-Levenshteinova vzdálenost

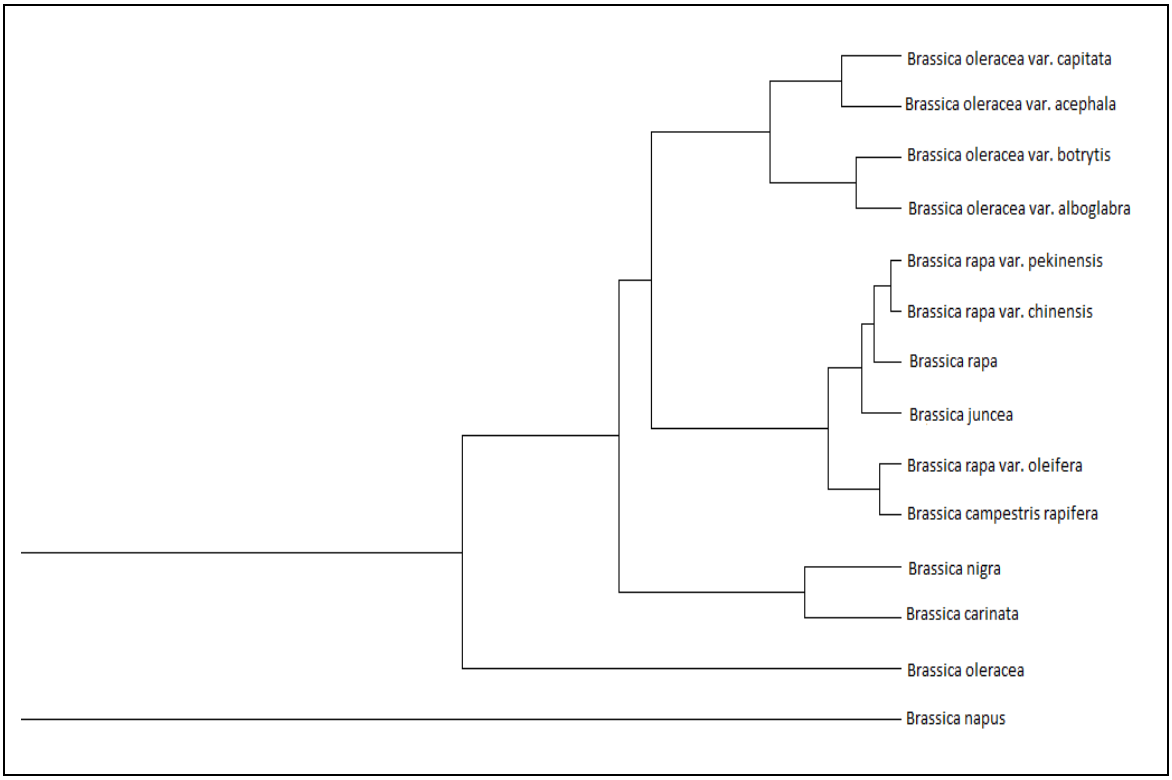
Jak bylo vysvětleno dříve, pro výpočet Damerau-Levenshteinovy vzdálenosti je potřeba zvolit si referenční sekvenci, se kterou jsou všechny ostatní ze zkoumaného vzorku porovnávány. Obvykle bývá při studiu rodu Brassica využíván jako referenční zmíněný druh Arabidopsis thaliana. Rozhodla jsem se, že pro tento výzkum zvolím jiný přístup a jako referenční vyberu tři druhy, konkrétně ty, které tvoří vrcholy trojúhelníku v U-modelu, tedy B. oleracea, B. rapa a B. nigra. Celkem tedy pro každý marker provedu tři výpočty Damerau-Levenshteinovy vzdálenosti, kdy každý bude reprezentovat, jak jsou ostatní druhy vzdáleny od jednoho z referenčních druhů. Důvodem pro volbu tohoto přístupu je předpoklad, že dosáhneme lepšího zmapování vzájemných vztahů v rámci rodu Brassica, pokud vezmeme v potaz podobnost druhů s každým ze svých progenitorů. Jak bylo popsáno výše, sledovanými změnami/transformacemi v této metodě jsou myšleny druhy bodových mutací, kterými se sekvence od sebe navzájem liší: *delece* (chybějící část), *inverze/substituce* (záměna částí), *inzerce* (začlenění části) nebo *transpozice* (prohození dvou sousedících částí). Pro každou z těchto transformací byla zvolena penalizace 1, pro shodu bází penalizace 0.

Výsledky analýzy markeru ITS jsou zaznamenány v tabulce (viz Tabulka 5), kde můžeme pozorovat, jak jsou jednotlivé druhy a jejich variety vzdáleny od jednoho z referenčních druhů. Vzhledem k charakteru metody se hodnoty proměňují podle toho, která sekvence je určena jako modelová a jsou podle ní vyhodnocovány změny. Pro vizualizaci takto zaznamenaných fylogenetických vztahů byla využita metoda hierarchického shlukování (viz Obrázek 19). V dendrogramu můžeme identifikovat pět větších větví:

1. B. oleracea var. capitata, B. oleracea var. acephala, B. oleracea var. botrytis, B. oleracea var. alboglabra
2. B. rapa var. pekinenesis, B. rapa var. chinensis, B. rapa, B. juncea, B. rapa var. oleifera, B. campestris rapifera
3. B. nigra, B. carinata
4. B. oleracea
5. B. napus

| | nigra | rapa | oleracea |
|---|---|---|---|
| acephala | 65 | 64 | 34 |
| alboglabra | 49 | 49 | 15 |
| botrytis | 43 | 40 | 20 |
| campestris | 53 | 14 | 48 |
| capitata | 62 | 52 | 24 |
| carinata | 23 | 46 | 53 |
| chinensis | 44 | 5 | 48 |
| juncea | 49 | 8 | 55 |
| napus | 222 | 226 | 201 |
| nigra | 0 | 58 | 56 |
| oleifera | 57 | 18 | 48 |
| oleracea | 100 | 88 | 0 |
| pekinensis | 46 | 3 | 48 |
| rapa | 49 | 0 | 48 |

**Tabulka 5**: Damerau-Levenshteinova vzdálenost markeru ITS1 5.8S ITS2



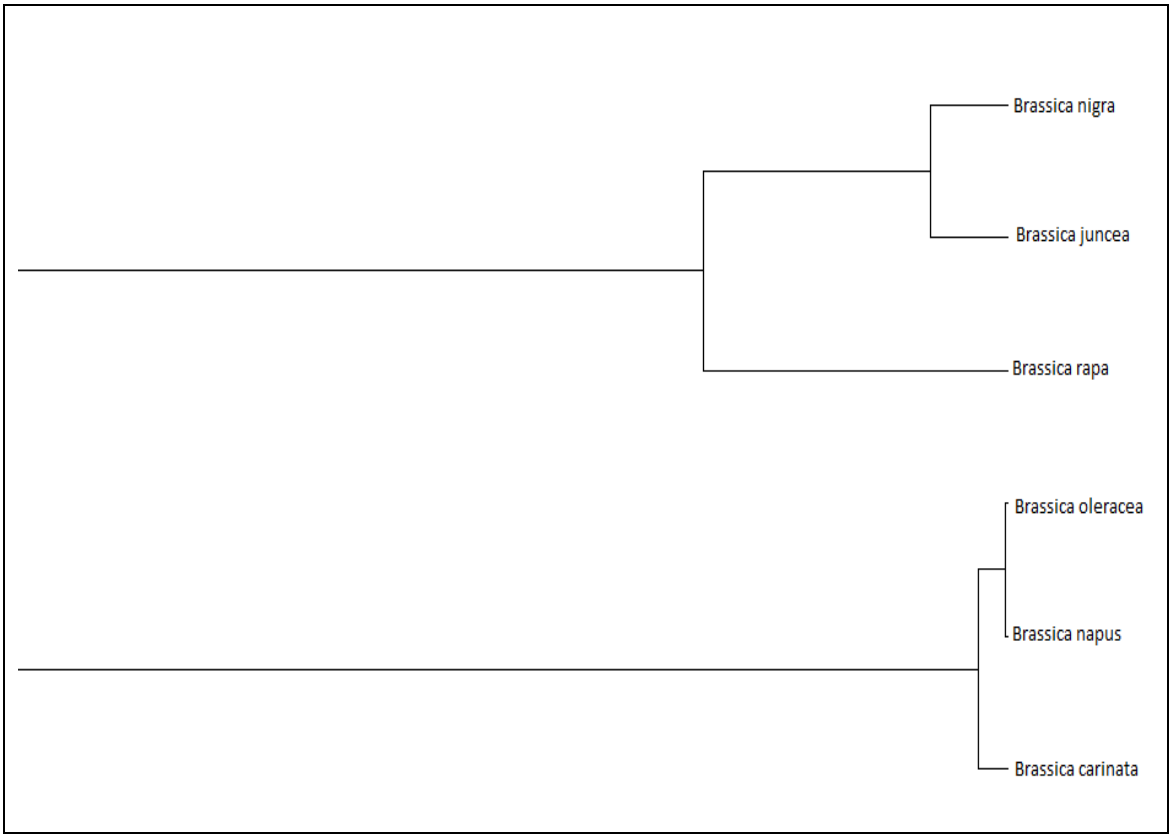

**Obrázek 19**: Dendrogram Damerau-Levenstheinovy vzdálenosti ITS markeru

V první větvi nacházíme shluk variet druhu B. oleracea, v druhé větvi nalézáme B. rapa spolu se svými varietami a také s druhem B. campestris rapifera a druhem B. juncea, pro který je B. rapa jedním z progenitorů. Ve třetí větvi se shlukuje B. nigra a B. carinata, mezi nimiž je také vztah progenitor-hybrid. Druh B. oleracea tvoří samostatnou větev a v garfu je umístěna mezi druhy B. carinata a B. napus, přičemž pro oba tyto druhy je jedním z progenitorů. Druh B. napus sice sousedí s B. oleracea, ale jinak je jeho větev zcela vyčleněna mimo ostatní druhy a variety. Také si můžeme povšimnout, že větev shlukující variety B. oleracea a větev B. rapa spolu s jejími varietami jsou ve vzájemné blízkosti, což odpovídá jejich blízkému genetickému vztahu. Můžeme tedy usoudit, že graf odpovídá sdříve popsaným fylogenetickým vztahům mezi druhy vzájemně a mezi druhy a jejich varietami. V případě hybridů došlo k jejich přičlenění k jednomu z progenitorů, z čehož můžeme usuzovat větší podíl genetické informace, která byla v případě tohoto markeru zděděna. Ačkoliv je B. napus v blízkosti svého progenitora B. oleracea, jeho celkové vyčlenění do samostatné větve dál od ostatních druhů může znamenat, že v případě tohoto druhu došlo k více změnám v procesu hybridizace. Tento předpoklad ověří analýza markeru matK a využití další metody.

Výsledky analýzy chloroplastového markeru matK (viz Tabulka 6) zaznamenávají pouze vztahy druhů tvořících trojúhelník U-modelu. Slouží tedy dpředevším k posouzení, zda zvolená analýza tohoto markeru dokáže identifikovat vzájemné vztahy progenitorů a hybridů. V grafu hierarchického shlukování (viz Obrázek 20) pak můžeme vidět pouze dvě hlavní větve – jedna je tvořena druhy B. nigra, B. juncea a B. rapa, druhá pak druhy B. oleracea, B. napus a B. carinata. Ve srovnání s dendrogramem ITS nedošlo v tomto případě k vydělení B. carinata a B. nigra do samostatné větve, místo toho se B. carinata přičlenila do větve k druhému ze svých progenitorů, B. oleracea, a B. nigra se přičlenila ke svému druhému hybridu, B. juncea. Zatímco v předchozím dendrogramu byly tyto druhy v samostatné větvia se svými dalšími příbuznými druhy pouze sousedily, zde došlo k rozpojení jejich původní vazby a připojení se do větve k těmto dříve sousedícím druhům – B. nigra k B. rapa a jejich hybridu B. juncea, B. carinata k progenitoru B. oleracea a jeho druhému hybridu B. napus. Můžeme tedy konstatovat, že analýza maturázy K potvrzuje vzájemné vztahy druhů tvořících U-model, ale některé mezidruhové vazby oslabuje, kdežto jiné posiluje.

| | oleracea | rapa | nigra |
|---|---|---|---|
| carinata | 34 | 1542 | 1179 |
| juncea | 14 | 358 | 77 |
| napus | 3 | 1525 | 1195 |
| nigra | 44 | 432 | 0 |
| oleracea | 0 | 1524 | 1197 |
| rapa | 16 | 0 | 80 |

**Tabulka 6**: Damerau-Levenshteinova vzdálenost markeru matK




**Obrázek 20**: Dendrogram Damerau-Levenshteinovy vzdálenosti matK

## Bag-of-Words model

Znovu připomeneme, že tzv. BoW model spočívá v reprezentaci textu jeho slovy, která jsou v případě genetických textů definována pomocí n-gramové analýzy. Zohledněna je frekvence slov a také multikódový charakter genetických sekvencí (jak jsem vysvětlila výše, předpokládáme-li, že genetické sekvence kódují různé funkce a v jejich sekvenci se jejich zápis překrývá, pak je n-gramová analýza vhodná k registraci těchto překryvů, protože okno výběru n-gramu postupuje po sekvenci vždy po jedné jednotce – aminokyselině/bázi). Také v analýze rodu Brassica byl využit program QUITA (*Quantitative Index Text Analyzer*), kde byly sekvence markerů tokenizovány (tj. rozčleněny na nukleotidové báze) a dále rozděleny na požadované n-gramy, tedy úseky stejné délky reprezentující slova. V tomto případě byly sledovány především změny, ke kterým dochází v zobrazení vztahů mezi jednotlivými druhy a jejich varietami a také mezi druhy a jejich hybridy, a to v závislosti na tom, zda genetický slovník markerů bude reprezentován slovy v podobě 3-gramů, 5-gramů a 10-gramů. Jak bylo dříve vysvětleno, změny je nutné sledovat, protože velikost gramů ovlivňuje podobu slov a jejich frekvenci. Pomocí tohoto postupu byl tedy marker ITS1 5.8S ITS2 i marker matK reprezentován přítomností či nepřítomností definovaných slov a jejich četností. Kvůli menšímu vzorku sekvencí pro marker matK jsou i zde výsledky jeho analýzy méně podrobné a poslouží zejména k zobrazení vztahů druhů tvořících U-model. Jelikož testované sekvence nemají stejnou délku, byla pro měření vzdálenosti mezi druhy použita cosinova distance. Pro vizualizaci podobnosti sekvencí bylo zvoleno hierarchické shlukování a MDS (*Multidimensional Scaling*).

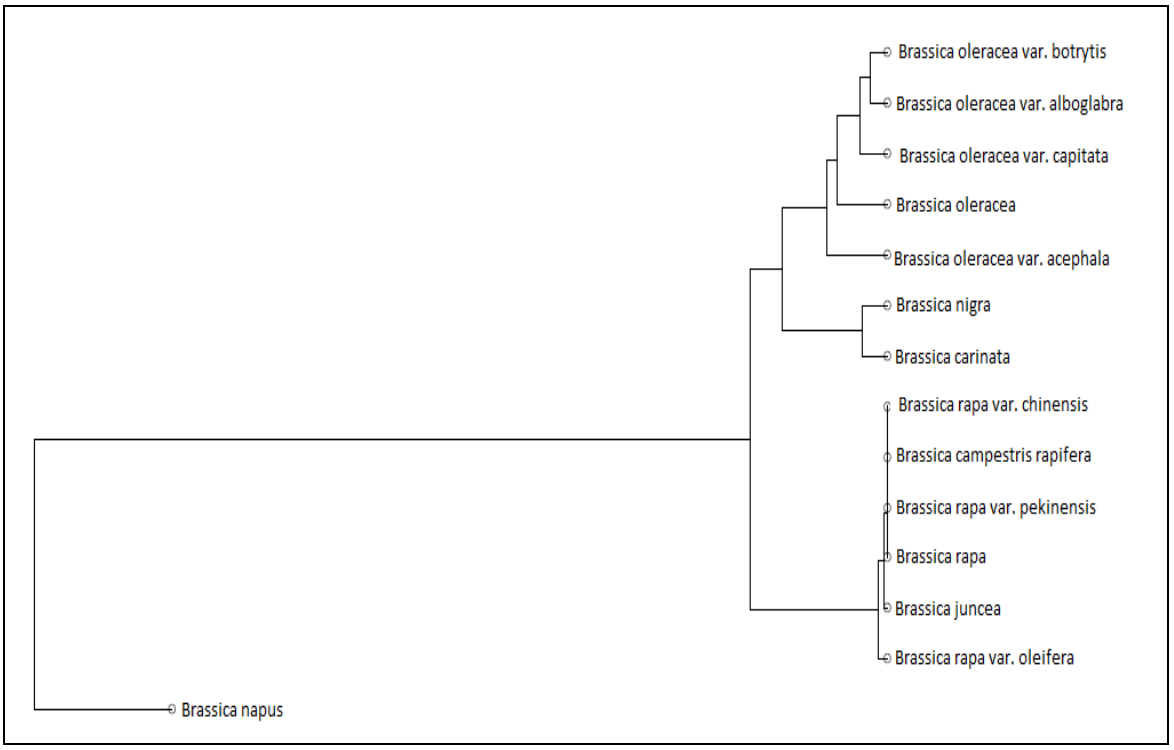
Když se podíváme na grafy hierarchického shlukování zobrazující fylogenetické vztahy na základě analýzy ITS markeru (viz Obrázek 21, Obrázek 23 a Obrázek 25), můžeme u všech identifikovat rozdělení druhů rodu Brassica do čtyř hlavních větví:

1. B. oleracea var. botrytis, B. oleracea var. alboglabra, B. oleracea var. capitata, B. oleracea, B. oleracea var. acephala
2. B. nigra, B. carinata
3. B. rapa var. chinensis, B. campestris rapifera, B. rapa var. pekinensis, B. rapa, B. juncea, B. rapa var. oleifera
4. B. napus

Povšimněme si, že rozdělení do větví je totožné jako u dendrogramu vytvořeného na základě Damerau-Levenshteinovy vzdálenosti ITS markeru, pouze s jediným rozdílem, a to tím, že se B. oleracea nevyčlenila do samostatné větve, ale nachází se v jedné větvi se svými varietami. Další změny v grafech se týkají podobnosti druhů v rámci dílčích větví a také vzájemné podobnosti mezi jednotlivými větvemi znázorněné délkou kladů.

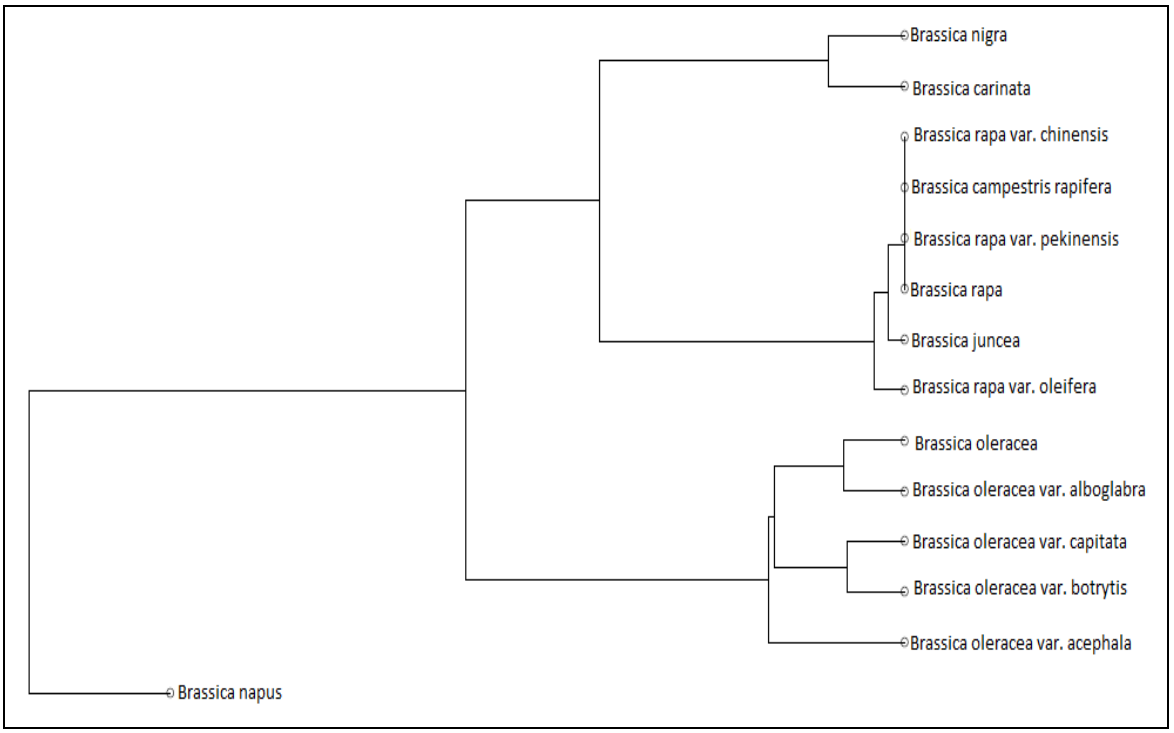
Porovnáme-li dendrogramy jednotlivých n-gramů mezi sebou, vidíme, že postupně dochází k prodlužování kladů, tedy ke zmenšování mezidruhové podobnosti. Tento jev byl dříve vysvětlen na základě změny slovníku, jímž jsou druhové sekvence reprezentovány – čím větší n-gram/specifičtější slovo, tím výraznější rozdílnosti jsou identifikovány. Z tohoto vývoje se vymyká pouze druh B. napus, který je ve stále stejné, a to výrazně vzdálené, pozici od ostatních větví. Jeho specifické postavení ukázala již dřívější analýza, zatím jej však nemůžeme vysvětlit. Co se týče ostatních větví, tak můžeme konstatovat, že odpovídají mezidruhovým fylogenetickým vztahům – B. oleracea je umístěna ve větvi se svými varietami, B. carinata se přičlenila ke svému progenitoru B. nigra. B. rapa se nechází ve větvi se svými varietami, dokonce můžeme podle kladů pozorovat naprostou shodu s B. rapa var. chinensis, B. rapa var. pekinensis a B. campestris rapifera. V případě B. rapa var. oleifera už byla detekována jistá odlišnost, stejně tak u hybridu B. juncea.

Prostorové zobrazení fylogenetických vztahů v grafech MDS (viz Obrázek 22, Obrázek 24 a Obrázek 26) není tak přehledné jako zobrazení v grafech hierarchického shlukování, ale lze u něj také zaznamenat proměny způsobené změnou velikosti n-gramů. Z velkého shluku viditelného v grafu 3-gramů se postupně (5-gramy -> 10-gramy) vydělil shluk, ve kterém je umístěna B. oleracea se svými varietami, a shluk druhu B. rapa s jejími varietami včetně B. carinata a B. juncea. V blízkosti tohoto shluku se vyskytuje B. nigra. Druh B. napus se ve všech grafech vyskytuje vzdálený od ostatních druhů, což koresponduje s jeho výrazným oddělením v dendrogramech. V grafech MDS tedy zobrazené vztahy korespondují s jejich zobrazením v grafech hierarchického shlukování, navíc je zde ještě viditelnější blízkost větve B. nigra a B. carinata s větví B. rapa a jejich variet.



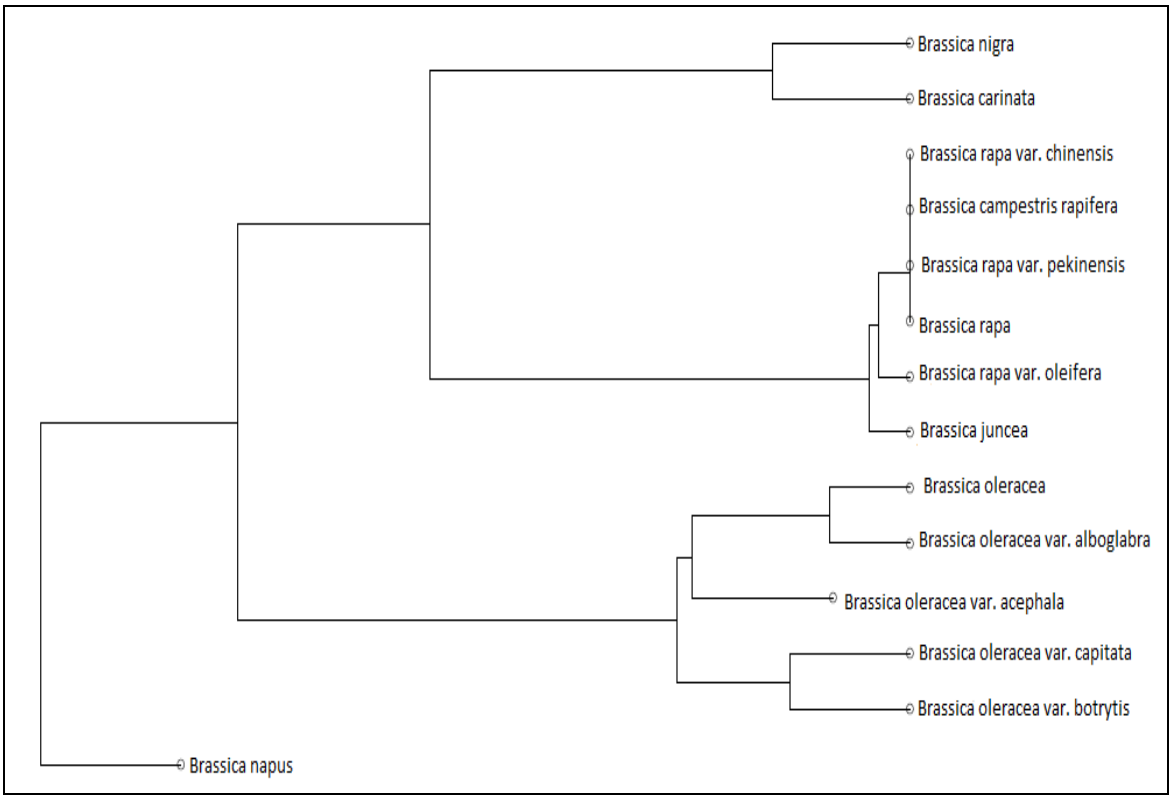

**Obrázek 21**: Dendrogram Bag-of-Words modelu ITS markeru, 3-gram


**Obrázek 22**: MDS graf Bag-of-Words modelu ITS markeru, 3-gram




**Obrázek 23**: Dendrogram Bag-of-Words modelu ITS markeru, 5-gram


**Obrázek 24**: MDS graf Bag-of-Words modelu ITS markeru, 5-gram

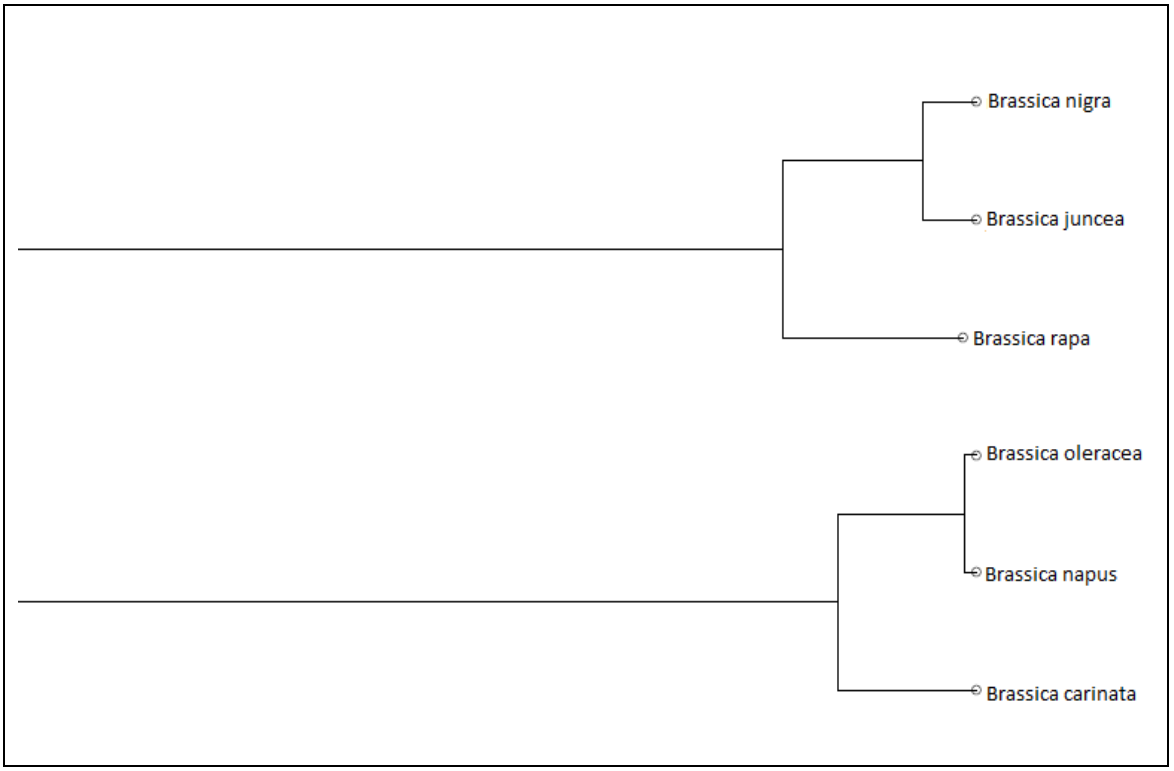



**Obrázek 25**: Dendrogram Bag-of-Words modelu ITS markeru, 10-gram

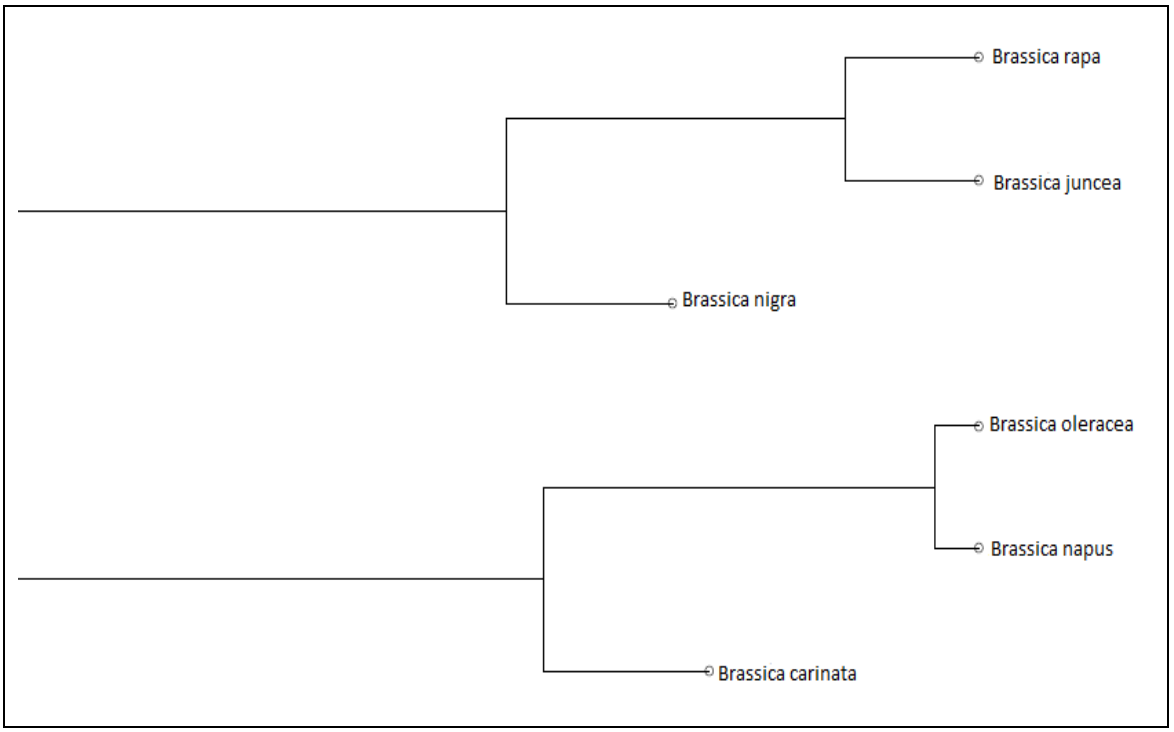

**Obrázek 26:** MDS graf Bag-of-Words modelu ITS markeru, 10-gram

Jelikož u markeru maturáza K pracujeme s velmi malým vzorkem a při analýze vlivem změny n-gramů nedošlo k výrazným změnám, jsou zde prezentovány pouze dendrogramy 3-gramů a 10-gramů (viz Obrázek 27 a Obrázek 28). Také u těchto grafů můžeme pozorovat stejné rozdělení druhů rodu Brassica do větví jako v grafu matK markeru u analýzy Damerau-Levenshteinovy vzdálenosti – jednu větev tvoří B. nigra, B. juncea a B. rapa, druhou větev tvoří B. olerace, B. napus a B. carinata.

Ve srovnání s dendrogramem ITS opět nedošlo k vydělení B. carinata a B. nigra do samostatné větve, ale ke shluku B. carinata se svým druhým progenitorem, B. oleracea, a B. nigra se přičlenila ke svému druhému hybridu, B. juncea. Je viditelné, že analýza maturázy K také na základě Bag-of-Words modelu potvrzuje vzájemné vztahy druhů tvořících U-model. Oproti markeru ITS však v obou analýzách zdůrazňuje vazby hybridů k jiným progenitorům. Také si lze povšimnout, že druh B. napus v této analýze nevykazuje takovou odlišnost od ostatních druhů a variet jako v analýze ITS. Což jen potvrzuje, že pro získání uceleného obrazu o fylogenetických vztazích je důležité v analýze využít více genetických markerů.




**Obrázek 27**: Dendrogram Bag-of-Words modelu matK markeru, 3-gram




**Obrázek 28**: Dendrogram Bag-of-Words modelu matK markeru, 10-gram

## Specifika hybridizace Barssicy napus

Analýza ITS markeru pomocí Damerau-Levenshteinovy vzdálenost i Bag-of-Words modelu ukázala, že druh Brassica napus vykazuje výraznější odlišnost od všech ostaních druhů i jejich variet. Ačkoliv je stejným typem hybridu jako Brassica juncea a Brassica carinata, tedy allotetraploidním hybridem (tj. obsahujícím čtyři genomy, odvozené ze dvou různých rodových druhů) a konkrétněji amfidiploidním (tj. obsahující jeden diploidní genom z každého ze dvou různých druhů Brassica), při vizualizaci vztahů pomocí clustrování se nepřičlenil ani k jednomu ze svých progenitorů – Brassica oleracea, Brassica rapa. Jelikož samotná vizualizace mezidruhových fylogenetických vztahů neumožní určit, proč je B. napus tak odlišný, je nutné nahlédnout dovnitř slovníku sekvencí, tedy do samotných n-gramů.

Pro detailnější vhled do podobnosti studovaných hybridů se svými progenitory byl vybrán frekvenční seznam jejich 7-gramů. Velikost n-gramů byla vybrána na základě studie ověřující platnost Zipfova zákona v genetických textech, která prokázala, že 7 bází genetického textu tvoří obdobný typ konstituentu jako pětice písmen přirozeného textu, tj. tvoří

obdobu slov (Faltýnek – Matlach 2014, s. 28–48). Jelikož byl výše pro analýzu genetických sekvencí využit Bag-of-Words model reprezentující text jeho slovy, je zde na místě porovnat podobnost příbuzných sekvencí na základě jejich analogických slov. Vzhledem k množství typů 7-gramů vyskytujících se v jedné sekvenci je pro demonstraci vztahů prezentován seznam pouze prvních patnácti typů; modrá barva v tabulkách značí stejnou frekvenci daného gramu u posuzovaných druhů. Jak je patrné (viz Tabulka 7), hybrid Brassica carinata zdědil genetickou informaci jaderné ribozomální DNA rovnoměrně od obou svých progenitorů, druhů Brassica oleracea a Brassica nigra.

| **CARINATA** | | **NIGRA** | | **OLERACEA** | |
|---|---|---|---|---|---|
| **Type** | **Freq** | **Type** | **Freq** | **Type** | **Freq** |
| a → c → t → c → t → c → g | 2 | a → c → t → c → t → c → g | 2 | g → g → a → t → a → t → c | 3 |
| a → g → a → a → c → g → a | 2 | a → g → a → a → c → g → a | 2 | a → a → a → t → c → g → t | 2 |
| c → g → a → t → g → a → a | 2 | c → a → a → g → c → c → t | 2 | a → a → t → c → g → t → c | 2 |
| c → g → t → c → c → c → c | 2 | c → g → a → t → g → a → a | 2 | a → c → t → c → t → c → g | 2 |
| c → t → c → t → c → g → g | 2 | c → g → t → c → c → c → c | 2 | c → a → a → a → t → c → g | 2 |
| g → c → c → g → a → t → t | 2 | c → t → c → t → c → g → a | 2 | c → g → g → a → a → g → c | 2 |
| g → t → g → a → a → t → t | 2 | c → t → c → t → c → g → g | 2 | c → g → g → a → t → a → t | 2 |
| t → c → c → c → g → t → g | 2 | g → c → c → g → a → t → t | 2 | c → g → g → t → t → g → g | 2 |
| t → c → t → c → g → g → c | 2 | g → c → t → c → t → c → g | 2 | c → t → c → t → c → g → g | 2 |
| t → c → t → c → g → g → t | 2 | g → t → g → a → a → t → t | 2 | g → t → g → a → a → t → t | 2 |
| t → t → c → c → g → t → g | 2 | t → c → c → c → g → t → g | 2 | g → t → t → t → c → g → g | 2 |
| a → a → a → a → c → g → a | 1 | t → c → g → a → t → g → a | 2 | t → c → c → c → g → t → g | 2 |
| a → a → a → a → g → c → t | 1 | t → c → g → g → t → c → g | 2 | t → c → t → c → g → g → c | 2 |
| a → a → a → a → g → t → g | 1 | t → c → t → c → g → g → c | 2 | t → t → t → c → g → g → t | 2 |
| a → a → a → a → t → c → c | 1 | t → c → t → c → g → g → t | 2 | a → a → a → a → c → g → a | 1 |

**Tabulka 7**: Podobnost B. carinata se svými progenitory B. oleracea a B. nigra, 7-gramy

**JUNCEA** **RAPA** **NIGRA**

| Type | Freq | Type | Freq | Type | Freq |
|---|---|---|---|---|---|
| a→c→t→c→t→c→g | 2 | a→c→t→c→t→c→g | 2 | a→c→t→c→t→c→g | 2 |
| a→g→a→a→c→g→a | 2 | a→g→a→a→c→g→a | 2 | a→g→a→a→c→g→a | 2 |
| c→g→a→t→g→a→a | 2 | c→g→a→t→g→a→a | 2 | c→a→a→g→c→c→t | 2 |
| c→g→g→a→a→g→c | 2 | c→g→g→a→a→g→c | 2 | c→g→a→t→g→a→a | 2 |
| c→g→g→a→t→a→t | 2 | c→g→g→a→t→a→t | 2 | c→g→t→c→c→c→c | 2 |
| c→t→c→t→c→g→g | 2 | c→t→c→t→c→g→g | 2 | c→t→c→t→c→g→a | 2 |
| g→c→c→t→g→c→t | 2 | g→c→c→t→g→c→t | 2 | c→t→c→t→c→g→g | 2 |
| g→g→a→t→a→t→c | 2 | g→g→a→t→a→t→c | 2 | g→c→c→g→a→t→t | 2 |
| g→t→g→a→a→t→t | 2 | g→t→g→a→a→t→t | 2 | g→c→t→c→t→c→g | 2 |
| t→c→c→c→g→t→g | 2 | t→c→c→c→g→t→g | 2 | g→t→g→a→a→t→t | 2 |
| t→c→t→c→g→g→c | 2 | t→c→t→c→g→g→c | 2 | t→c→c→c→g→t→g | 2 |
| t→g→g→c→c→a→a | 2 | t→g→g→c→c→a→a | 2 | t→c→g→a→t→g→a | 2 |
| t→t→g→g→c→c→a | 2 | t→t→g→g→c→c→a | 2 | t→c→g→g→t→c→g | 2 |
| a→a→a→a→c→g→a | 1 | a→a→a→a→c→g→a | 1 | t→c→t→c→g→g→c | 2 |
| a→a→a→a→g→c→t | 1 | a→a→a→a→g→c→t | 1 | t→c→t→c→g→g→t | 2 |

**Tabulka 8**: Podobnost B. juncea se svými progenitory B. rapa a B. nigra, 7-gramy

**NAPUS** **OLERACEA** **RAPA**

| Type | Freq | Type | Freq | Type | Freq |
|---|---|---|---|---|---|
| g→g→g→t→g→c→g | 4 | g→g→a→t→a→t→c | 3 | a→c→t→c→t→c→g | 2 |
| c→g→c→c→g→c→g | 3 | a→a→a→t→c→g→t | 2 | a→g→a→a→c→g→a | 2 |
| g→c→c→g→c→g→g | 3 | a→a→t→c→g→t→c | 2 | c→g→a→t→g→a→a | 2 |
| g→c→t→c→c→c→g | 3 | a→c→t→c→t→c→g | 2 | c→g→g→a→a→g→c | 2 |
| g→g→g→g→t→g→c | 3 | c→a→a→a→t→c→g | 2 | c→g→g→a→t→a→t | 2 |
| c→c→a→c→c→c→c | 2 | c→g→g→a→a→g→c | 2 | c→t→c→t→c→g→g | 2 |
| c→c→c→a→c→c→c | 2 | c→g→g→a→t→a→t | 2 | g→c→c→t→g→c→t | 2 |
| c→c→c→c→a→c→c | 2 | c→g→g→t→t→g→g | 2 | g→g→a→t→a→t→c | 2 |
| c→c→g→c→g→g→c | 2 | c→t→c→t→c→g→g | 2 | g→t→g→a→a→t→t | 2 |
| c→c→g→c→g→g→g | 2 | g→t→g→a→a→t→t | 2 | t→c→c→c→g→t→g | 2 |
| c→g→c→c→c→c→a | 2 | g→t→t→t→c→g→g | 2 | t→c→t→c→g→g→c | 2 |
| c→g→c→g→g→g→g | 2 | t→c→c→c→g→t→g | 2 | t→g→g→c→c→a→a | 2 |
| c→g→c→t→c→c→c | 2 | t→c→t→c→g→g→c | 2 | t→t→g→g→c→c→a | 2 |
| c→g→c→t→c→g→c | 2 | t→t→t→c→g→g→t | 2 | a→a→a→a→c→g→a | 1 |
| c→g→g→g→g→t→g | 2 | a→a→a→a→c→g→a | 1 | a→a→a→a→g→c→t | 1 |

**Tabulka 9**: Podobnost B. napus se svými progenitory B. oleracea a B. rapa, 7-gramy

Hybrid Brassica juncea (viz Tabulka 8) vykazuje naprostou shodu se svým progenitorem Brassica rapa a částečnou podobnost s progenitorem Brassica nigra, genetická infromace jaderné ribozomální DNA byla tedy dominantně převzata od B. rapa. Zcela odlišnou situaci však můžeme pozorovat u hybridu Brassica napus (viz Tabulka 9), který nevykazuje jakoukoliv podobnost se svými progenitory, druhy Brassica oleracea a Brassica rapa. Modře zvýrazněná pole v tabulce značí podobnost mezi oběma progenitory, která odpovídá hypotéze, že tyto druhy si jsou fylogeneticky velmi blízké (Zielkowski – Lysak – Heneen 2011, s. 257–265). Tato porovnání frekvenčních pořadí typů 7-gramů potvrzují, že u druhu B. napus došlo k jinému typu hybridizace než u druhů B. carinata a B. juncea, které jednoznačně vykazují podobnost se svými progenitory. Je ovšem nemožné pouze na základě provedených analýz tento jev vysvětlit. Možným vysvětlením je, že u tohoto druhu došlo k několikanásobné hybridizaci a podobnost s původními progenitory již není v jaderné ribozomální DNA patrná. Odvážnější by bylo domnívat se, že skutečnými progenitory jsou jiné druhy, které mají stejný počet chromozomů jako druhy B. oleracea a B. rapa. Pro ověření těchto předpokladů je však nutné podrobit analýze ostatní zástupce rodu Brassica, které netvoří U-model, případně analyzovat druhy zařazené do tribu Brassiceae.

## Závěr

Cílem této kapitoly bylo v návaznosti na výzkumy týmu Martina Lysáka i zahraniční studie týkající se fylogenetických vztahů v rámci rodu Brassica otestovat potenciál lingvistických metod mapovat příbuzenské vztahy druhů tvořících tzv. U-model – jedná se o příbuzenské vztahy mezidruhové, mezi druhy a jejich varietami, mezi druhy a jejich hybridy. K tomu byly využity metody Damerau-Levenshteinova vzdálenost a Bag-of-Words model, které se osvědčily při předešlé studii rodu Malus. Testovány byly molekulární markery, konkrétně chloroplastový marker maturáza K a jaderný ribozomální marker ITS1 5.8S ITS2. Obě metody se ukázaly být vhodnými pro mapování fylogenetických vztahů na mezidruhové úrovni, včetně vztahů k jejich varietám. Co se týče vztahů mezi druhy a jejich hybridy, bylo detekováno specifické postavení hybridu Brassica napus, který v žádné analýze nevykazoval podobnost se svými progenitory, což bylo potvrzeno detailnějším porovnáním frekvenčního seznamu 7-gramů jejich ITS markerů. Vysvětlení tohoto jevu však vyžaduje studium dalších rodových či tribových druhů, jež může poskytnout komplexnější pohled na příbuznost druhu B. napus k ostatním příbuzným druhům a lépe tak zmapovat charakter jeho hybridizace.

## Diskuze

Ačkoliv užité lingvistické metody prokázaly potenciál uplatnit se v molekulárně fylogenetických studiích a obohatit svými výsledky standardně užívané bioinformatické postupy, je na místě zmínit několik poznatků o problematických aspektech jejich využití:

- V prezentovaných studiích byly použity pouze dva typy markerů – chloroplastový marker maturáza K a jaderný ribozomální marker ITS. Tím nelze zevšeobecnit využitelnost těchto metod na libovolném genetickém materiálu. Šíře jejich využití musí být ověřena provedením dalších analýz na jiných genetických markerech, obecněji na genetickém materiálu s různými vlastnostmi.
- Ve všech analýzách byl pro každý druh či varietu použit pouze jeden vzorek. To umožnilo vytvořit reprezentativní vizualizaci fylogenetických vztahů. Použití více vzorků pro každý zkoumaný druh by nemuselo na výsledcích nic změnit, ale také by mohla být detekována variabilita markeru v rámci jednoho druhu nebo složitější mezidruhové vztahy.
- Testovaný jaderný ribozomální marker i chloroplastový marker byly zapsány v nukleových bázích. Vzhledem k výzkumnému cíli zachytit co nejpřesněji rozdíly mezi druhovými sekvencemi vyvstala potřeba analyzovat také proteiny v podobě zápisu bází. Tímto způsobem je možné získat o rozdílech v genetickém zápisu jednotlivých sekvencí co nejvíc informací, a to na úrovni bodových mutací.
- Užité metody slouží k detekování vzdálenosti/míry podobnosti mezi sekvencemi a sestavení dendrogramu. Charakter vztahů jednotlivých druhů však není zcela odvoditelný z pouhého grafu. Proto je nutné zohlednit např. agrikulturní využití zkoumaných rostlin, případně modifikovat metodologii např. zařazením vzorků dalších příbuzných druhů do analýzy, využitím další metody analýzy, analyzováním jiného genetického materiálu – specifická situace byla patrná u hybridu Brassica napus.
- V této práci byly lingvistické metody využity k testování již existujících hypotéz, které byly formulovány na základě standardních bioinformatických analýz. Otázkou zůstává, zda je potenciál těchto metod natolik silný, aby výsledky na základě nich provedených analýz mohly být východiskem pro nové hypotézy. Nejvhodnějším způsobem, jak tento potenciál podpořit, je testovat co největší vzorek druhových sekvencí a využít pro analýzu více různých molekulárních znaků.

# Závěr

Cílem této práce bylo představit možnosti využití lingvistických metod (*linguistics-like tools*) v genetice a blíže pak v molekulrání fylogenetice. Pozornost byla zaměřena na Damerau-Levenshteinovu vzdálenost a Bag-of-Words model. Potenciál těchto metod byl využit k testování dvou hypotéz. V první analýze se jednalo o zjištění původu domácí jabloně (Malus domestica). Byla potvrzena hypotéza o její přímé fylogenetické vazbě na planou jabloň Sieversovu (Malus sieversii) pocházející z oblasti Ťian Šanu v Centrální Asii. Lingvistická analýza tak přinesla další důkaz vyvracející dřívější předpoklad, že domácí jabloň vznikla zkřížením několika odrůd. Druhá studie byla zaměřena na mapování vztahů na mezidruhové úrovni, vztahů také mezi druhy a jejich varietami a mezi druhy a jejich hybridy tvořícími tzv. U-model rodu Brassica (brukev). V tomto případě se ukázalo, že na základě výsledků zvolených metod je možné vytvořit dendrogram odrážející komplexní fylogenetické vztahy na nižších taxonomických úrovních. Analýzy také odhalily odlišný charakter hybridizace u druhu brukev řepka (Brassica napus), který se ve všech analýzách vyčleňoval od ostatních rodově příbuzných druhů a nevykazoval podobnost se svými předpokládanými progenitory (Brassica oleracea, Brassica rapa).

Lze tedy říct, že zvolené lingvistické metody jsou vhodným nástrojem analýzy genetického materiálu pro fylogenetické studie. V diskuzi byly jmenovány aspekty, které je třeba při jejich využití vzít v úvahu – charakter genetického materiálu (kódující DNA, nekódující DNA, RNA, protein atd.), počet analyzovaných vzorků, zápis sekvencí v aminokyselinách či bázích, širší znalosti o studovaném materiálu pro detekování povahy odlišností mezi jednotlivými druhy, potenciál metod poskytnout silná data pro budování nových hypotéz. Jak bylo uvedeno, validitu výsledků lze posílit testováním co největšího vzorku sekvencí a využitím více druhů genetických markerů. Sledování fylogenetických změn je pak nejvhodnější při práci se sekvencemi zapsaných v nukleových bázích, protože tak nedochází ke ztrátě informace (jak by tomu bylo u aminokyselin) a lze sledovat konkrétní změny na úrovni bodových mutací.

Dalším krokem je ověřit využitelnost testovaných i dalších lingvistických metod na jiném biologickém materiálu (z rostlinné i živočišné říše) a také na jiných molekulárních znacích. Tím by bylo posíleno postavení těchto metod vedle standardních bioinformatických metod užívaných v molekulární fylogenetice.

# Přílohy

## Seznam použitých sekvencí k analýze rodu Malus

Zdroj: https://www.ncbi.nlm.nih.gov/genbank/

| | **ITS1 5.8S ITS2** | **matK** |
|---|---|---|
| Malus asiatica | EF442030.1 | AF309174.1 |
| Malus baccata | EF525562.1 | AF309178.1 |
| Malus coronaria | AF186524.1 | AF309186.1 |
| Malus domestica-Ashmead´s Kernal | AF186480.1 | AF309171.1 |
| Malus domestica-Bramley´s Seedling | AF186479.1 | AF309172.1 |
| Malus doumeri | AF186529.1 | AF309191.1 |
| Malus florentina | AF186520.1 | AF309182.1 |
| Malus fusca | AF186514.1 | AF309183.1 |
| Malus halliana | AF186502.1 | KP089151.1 |
| Malus hupehensis | AF186503.1 | AF309179.1 |
| Malus ioensis | AF186526.1 | AF309187.1 |
| Malus kansuensis | AF186512.1 | AF309193.1 |
| Malus micromalus | EF525565.1 | AF309175.1 |
| Malus prattii | AF186511.1 | AF309188.1 |
| Malus prunifolia | AF186500.1 | JQ391019.1 |
| Malus sieversii | AF186485.1 | AF309173.1 |
| Malus sylvestris | JQ392462.1 | AF309177.1 |
| Malus toringoides | AF186517.1 | AF309180.1 |
| Malus trilobita | AF186521.1 | AF309189.1 |
| Malus tschonoskii | AF186527.1 | AF309189.1 |
| Malus yunnanensis | AF186508.1 | AF309192.1 |

## Seznam použitých sekvencí k analýze rodu Brassica

Zdroj: https://www.ncbi.nlm.nih.gov/genbank/

| | **ITS1 5.8S ITS2** | **matK** |
|---|---|---|
| Brassica carinata | DQ003690.1 | AB354275.1 |
| Brassica campestris rapifera | GQ268060.1 | |
| Brassica juncea | AF128093.1 | AB354274.1 |
| Brassica oleracea | AY722423.1 | AB354271.1 |
| Brassica oleracea var. acephala | GQ891869.1 | |
| Brassica oleracea var. alboglabra | GQ891870.1 | |
| Brassica oleracea var. botrytis | GQ891875.1 | |
| Brassica oleracea var. capitata | DQ003650.1 | |
| Brassica napus | AB456109.1 | AB354273.1 |
| Brassica nigra | DQ003644.1 | AB354272.1 |
| Brassica rapa | AF531563.1 | AY541619 |
| Brassica rapa var. chinensis | AF128095.1 | |
| Brassica rapa var. oleifera | GQ891873.1 | |
| Brassica rapa var. pekinensis | AF128096.1 | |

# Bibliografie


Al-Shehbaz, A. I. – Beilstein, M. A. – Kellogg, E. A. (2006). Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Syst. Evol*. Vol. 259, s. 89–120.

Barbieri, M. (2006). *Organické kódy*. Praha: Academia.

Barbieri, M. (2008a). Life is Semiosis. The biosemiotic view of Nature. *Cosmos and History: The Journal of Natural and Social Philosophy*, Vol. 4, s. 29–52.

Barbieri, M. (2008b). *A Critique of Biohermeneutics*.

Barbieri, M. (2009). A Short History of Biosemiotics. *Biosemiotics* . Vol. 2, Issue 2, s. 221–245.

Barbieri, M. (2015). *Code Biology. A New Science of Life*. Dordrecht: Springer.

Beadle, G. – Beadle, M. (1966). *The Language of Life. An Introduction to the Science of Genetics.* New York: Doubleday and Co.

Blažek, J. et al. (1998). *Ovocnictví*. Praha: Květ.

Bolshoy, A. (2003). DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity. *Applied bioinformatics*. Vol. 2. s. 103–112.

Borkhausen, M. B. (1803). *Theoretisch-praktisches Handbuch der Forstbotanik und Forsttechnologie*. Wien: Georg Friedrich Heyer.

Brickell, Ch. (2003). *A–Z Encyclopedia of garden plants*. The Royal Horticultural Society.

Cornille, A. et al. (2012). New Insight into the History of Domesticated Apple: Secondary Contribution of the European Wild Apple to the Genome of Cultivated Varieties. *PLoS Genetics*, Vol. 8(5), e1002703.

Cvrčková, F. (2006). *Úvod do praktické bioinformatiky*. Praha: Academia.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors, *Communications of the ACM*, Vol. **7** (3): s. 171–176.

Eriksson, T. et al. (2003). The Phylogeny of Rosoideae (Rosaceae) Based on Sequences of the Internal Transcribed Spacers (ITS) of Nuclear Ribosomal DNA and the trnL/F Region of Chloroplast DNA. *International Journal of Plant Sciences*, Vol. 164:2, s. 197–211.

Faltýnek, D. – Matlach, V. (2014). *Gramatiky DNA*. Olomouc: Univerzita Palackého.

Filliat, D. (2007). A visual bag of words method for interactive qualitative localization and mapping. *International Conference on Robotics and Automation*, s. 3921–3926.

Flegr, J. (2005). *Evoluční biologie*. Praha: Academia.

Forsline, P. L. et al. (1994). Collection of wild Malus, Vitis and other fruit species genetic resources in Kazakstan and neighboring republics. *HortScience*, Vol. 29, no. 433.

Franzke, A. et al. (2011). Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends in Plant Science*, Vol. 16(2), s. 108–116.

Gharghani, A. et al. (2009). Genetic identity and relationships of Iranian apple (Malus × domestica Borkh.) cultivars and landraces, wild Malus species and representative old apple cultivars based on simple sequence repeat (SSR) marker analysis. *Genet Resour Crop Evol*, Vol. 56, s. 829–842.

Gross, B. L. et al. (2014). Genetic diversity in Malus ×domestica (Rosaceae) through time in response to domestication. *Am J Bot*. Vol. 101(10), s. 1770–1779.

Harris, S. A. – Robinson, J. P. – Juniper, B. E. (2002). Genetic clues to the origin of the apple. *Trends in Genetics*, Vol. 18(8), s. 426–430.

Harris, Z. (1954). Distributional Structure. *Word*. Vol. 10 (2/3), s 146–162.

Hokanson, S. – Lamboy, W. – Szewc-McFadden, A. – McFerson, J. (2001). Microsatellite (SSR) variation in a collection of Malus (apple) species and hybrids. *Euphytica*, Vol. 118, s. 281–294.

Cheng, F. – Lysak, M. A. – Mandáková, T. – Wang, X. (2015). The common ancestral genome of the *Brassica* species. In: Wang, X. – Kole, Ch. (eds.): *The Brassica rapa Genome*. USA: Springer, s. 97–106.

Jakobson, R. (1971). *Selected Writings II: Word and Language*. The Hague: Mouton & Co.

Jakobson, R. (1973). La linguistique et les sciences naturelles. In *Essais de linguistique générale 2: Rapports internes et externes du langage*. Paris: Les Éditions de Minuit.

Jakobson, R. (1974). Life and Language. *Linguistics*, Vol. 138, s. 97–103.

Janick, J. – Moore, J. N. (1996). Apples. In *Fruit Breeding: Tree and Tropical Fruits*, s. 1–77, New York: John Wiley & Sons.

Ji, S. (1991). Biocybernetics: a machine theory of biology. In: *Molecular Theories of Cell Life and Death*. New Brunswick: Rutgers University Press.

Ji, S. (1997). Isomorphism between cell and human languages: molecular biological, bioinformatic and linguistic implications. *Biosystems*. Vol. 44, Issue 1, s. 17–39.

Ji, S. (1999). The Linguistics of DNA: Words, Sentences, Grammar, Phonetics, and Semantics. *Annals of the New York Academy of Sciences*, Vol. 870, s. 411–417.

Jonák, J. (2007). RNA v proteosyntéze: Genetický kód a příprava aminoacyl-tRNA. *Živa*, Vol. 5. s. 195–198.

Juniper, B. E. – Watkins, R. – Harris, S. A. (1998). The origin of the apple. *Acta Horticulturae*, Vol. 484. s. 27–33.

Kiefer, M. – Mandáková, T. – Lysak, M. A. (2014). BrassiBase: Introduction to a novel knowledge database on Brassicaceae evolution. *Plant & Cell Physiology*, Vol. 55(1): e3.

Koidzumi, G. (1934). Contributiones ad floram asiae orientalis. *Acta Phytotaxonomica et Geobotanica*, Vol. 3, s. 146–155.

Kull, K. (2007): A Brief History of Biosemiotics. In: Barbieri, M. (Ed.). *Biosemiotics: Information, Codes and Signs in Living Systems* (1−26). New York: Nova Science Publisher Inc.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*, Vol. 10 (8), s. 707–710.

Linnæi, C. *(Carolus Linnæus)* (1758). *Systema naturæ per regna tria naturae: secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis* (10. vyd.). Holmiæ (Stockholm): Laurentius Salvius.

Lovato, P. (2015). *Bag of words approaches for Bioinformatic*, Ph.D. thesis, Dept. of Computer Science, University of Verona, series TD-03-15.

Lysák, M. A. (2009). Comparative cytogenetics of wild crucifers (Brassicaceae). In: Gupta SK (ed.): *Wild Crucifers – Biology, Breeding and Utilization*. Boca Raton: Taylor and Francis Group, FL, s. 177–205.

Lysák, M. A. – Cheung, K. – Kitschke, M. – Bu, P. (2007). Ancestral Chromosomal Blocks Are Triplicated in Brassiceae Species with Varying Chromosome Number and Genome Size. *Plant Physiology*. Vol. 145 (2), s. 402–410.

Lysák, M. A. – Koch, M. A. (2011). Phylogeny, genome and karyotype evolution of crucifers (Brassicaceae). In: Bancroft, I. – Schmidt, R. (eds.): *Genetics and Genomics of the Brassicaceae*. New York: Springer. s. 1–31.

Majorek, A. K. (2014). The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification, *Nucleic Acids Research*, Vol. 42, s. 4160–4179.

Mantegna, R. N. et al. (1995). Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, Vol. 52, s. 2939–2950.

Markoš, A. (2003). *Tajemství hladiny. Hermeneutika živého*. Praha: Dokořán.

Markoš, A. a kol. (2014). *Biosémiotika 2*. Olomouc: Univerzita Palackého v Olomouci.

Ohno, S. (1992). Of palindromes and peptides. *Genomics*, Vol. 90, s. 342–345.

Page, R. D. M. – Holmes, E. C. (1998). *Molecular evolution: a phylogenetic approach*. Oxford: Blackwell Science.

Parkin, A. P. I. et al. (2005). Segmental Structure of the *Brassica napus* Genome Based on Comparative Analysis With Arabidopsis thaliana. *GENETICS*, Vol. 171, no. 2, s. 765–781.

Phipps, J. B. – Robertson, K. R. – Smith, P. G. – Rohrer, J. R. (1990). A checklist of the subfamily Maloideae (Rosaceae). *Canadian Journal of Botany*, 68, s. 2209–2269.

Popov, O. – Segal, D. M. – Trifonov, E. N. (1996). Linguistic complexity of protein sequences as compared to texts of human languages. *Biosystems*, Vol. 38, Issue 1, s. 65–74.

Rani, M. – Mitra, C. K. (1994). Periodicities in protein sequences. *Journal of Biosciences*, Vol. 19 (2), s. 255–266.

Rehder, A. (1949). *Bibliography of cultivated trees and shrubs.* Jamaica Plain: The Arnold Arboretum of Harvard University.

Royo, B. J. – Itoiz, R. (2004). Evaluation of the discriminance capacity of RAPD, isoenzymes and morphologic markers in apple (*Malus x domestica* Borkh.) and the congruence among classifications. *Genetic Resources and Crop Evolution*. Vol. 51, s. 153–160.

Sanders**,** N. C. – Chin, S. B. (2009). Phonological Distance Measures. *Journal of Quantitative Linguistics*, Vol. *16*(1), s. 96–114.

Serva, M. – Petroni, I. F. (2007). Indo-European Languages Tree by Levenshtein Distance. *EPL (Europhysics Letters)*, Vol. 81, s. 680–685.

Schmidt, R. – Banncroft, I. (2011). *Genetics and genomics of the Brassicaceae*. New York: Springer.

Schulz, M. A. – Schmalbach, B. – Brugger, P. – Witt, K. (2012). Analysing Humanly Generated Random Number Sequences: A Pattern-Based Approach. *PLoS ONE*, Vol. 7(7), e41531.

Sikka, K. – Wu, T. – Susskind, J. – Bartlett, M. (2012). Exploring Bag of Words Architectures in the Facial Expression Domain. In: Fusiello, A. – Murino, V. – Cucchiara, R. (eds) Computer Vision – ECCV 2012. Workshops and Demonstrations. ECCV 2012. *Lecture Notes in Computer Science*, Vol. 7584. Berlin: Springer.

Sutton, L. A. et al. (2014). An Entity Evolving into a Community: Defining the Common Ancestor and Evolutionary Trajectory of Chronic Lymphocytic Leukemia Stereotyped Subset #4. *Molecular Medicine*, Vol. 20(1), s. 720–728.

Toldo, R. – Castellani, U. – Fusiello, A. (2009). A Bag of Words Approach for 3D Object Categorization. In: Gagalowicz, A. – Philips, W. (eds) Computer Vision/Computer Graphics CollaborationTechniques. MIRAGE 2009. *Lecture Notes in Computer Science*, Vol. 5496. Berlin: Springer.

Town, C. D. et al. (2006). Comparative Genomics of Brassica oleracea and Arabidopsis thaliana Reveal Gene Loss, Fragmentation, and Dispersal after Polyploidy. *The Plant Cell*, Vol. 18(6), s. 1348–1359.

Trifonov, E. N. (1989). The multiple codes of nucleotide sequences. *Bulletin of Mathematical Biology*. Vol. 51 (4), s. 417–432.

Trifonov, E. N. (1990). Making sense of the human genome. *Structure and Methods: Proceedings of the Sixth Conversation in the Discipline Biomolecular Stereodynamics*. New York: Adenine Press. Vol. 1, s. 69–78.

U, N., (1935). Genome analysis in Brassica with special reference to the experimental formation ofB. napus and peculiar mode of fertilization. *Jap. J. Bot*. Vol. 7, s. 389–452.

Vavilov, N.I. (1930). Wild progenitors of the fruit trees of Turkistan and the Caucasus and the problém of the origin of fruit trees. *International Horticultural Congress Group B*, s. 271–286.

Wang, J. et al. (2013). Bag-of-words representation for biomedical time series classification, *Biomedical Signal Processing and Control*, Vol. 8, Issue 6, s. 634–644.

Yamamoto, M. – Nishio, T. (2014). Commonalities and differences between Brassica and Arabidopsis self-incompatibility. *Horticulture Research*, Vol. 1, article num. 1.

Zemková, M. (2016). *Lingvistické přístupy v genomice a lingvistická metafora v biologii.* Disertační práce. Katedra filosofie a dějin přírodních věd. Přírodovědecká fakulta. Univerzita Karlova v Praze.

Zhang, Y. et al. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*. Vol. 1, Issue 1–4, s. 43–52.

Ziolkowski, P. A. – Lysak, M. A. – Heneen, W. K. (2011). Cytogenetic Studies in Vegetable Brassicas. s. 257–303. In Sadowski, J. – Kole, C. (2011). *Genetics, Genomics and Breeding of Vegetable Brassicas*. Enfield, NH, USA: Science Publishers.