VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

NUMERICKÉ REPREZENTACE PROTEINOVÝCH SEKVENCÍ PRO KLASIFIKACI
NUMERICAL REPRESENTATIONS OF PROTEIN SEQUENCES FOR CLASSIFICATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Vojtěch Bartoň**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Denisa Maděránková**
SUPERVISOR

BRNO 2015

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Biomedicínská technika a bioinformatika**
Ústav biomedicínského inženýrství

Student: Vojtěch Bartoň ***ID:*** 164962
Ročník: 3 ***Akademický rok:*** 2015/16

NÁZEV TÉMATU:

Numerické reprezentace proteinových sekvencí pro klasifikaci

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma metody numerické reprezentace proteinových sekvencí. 2) V libovolném programovém prostředí naprogramujte alespoň tři vybrané numerické reprezentace. 3) Navrhněte a naprogramujte metodiku porovnávání naprogramovaných reprezentací pro účely fylogenetické klasifikace. 4) Naprogramované funkce otestujte na setu jaderně a mitochondriálně kódovaných proteinů. 5) Výsledky vyhodnoťte, diskutujte a porovnejte se standardní taxonomií.

DOPORUČENÁ LITERATURA:

[1] ZHANG, Y.-P., et al. Novel Numerical Characterization of Protein Sequences Based on Individual Amino Acid and Its Application. BioMed Research International. 2015, ID 909567.

[2] DENG, W. a LUAN, Y. DV-Curve Representation of Protein Sequences and Its Application. Computational and Mathematical Methods in Medicine. 2014, 203871.

Termín zadání: 8.2.2016 ***Termín odevzdání:*** 27.5.2016

Vedoucí práce: Ing. Denisa Maděránková
Konzultant bakalářské práce:

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

UPOZORNĚNÍ:
Autor bakalářské práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

Abstrakt

S rozmachem bioinformatiky vyvstala možnost analyzovat a srovnávat i rozsáhlé soubory nejen genomických, ale i proteomických sekvencí. Byla tedy nutnost zavést numerické reprezentace sekvencí pro jejich počítačové zpracování. Reprezentace proteinových sekvencí má svá specifika a často vyšší výpočetní náročnost, než reprezentace genomických sekvencí. V práci je představeno několik různých metod přístupu k numerickým reprezentacím proteinů. Vybrané metody jsou poté testovány na setu mitochondriálně kódovaných proteinů a srovnány se standardní taxonomií i s běžně používanou symbolickou reprezentací.

Klíčová slova

Numerické reprezentace proteinů, proteinové sekvence, mitochondriální proteiny, klasifikace organismů, bioinformatika

Abstract

Todays we have the opportunity to analyze huge sets of genomics and proteomics data. In my bachaleor thesis I introduce a few numerical alternatives to represent proteins. The usage of numerical representations opened the way to analyze proteomics data as digital signals, which bring us quantity of new possibilities how to process the protein. In my thesis I compare a few numerical representation with standard taxonomy and with symbolic representation too.

Keywords

Numerical protein representation, protein sequences, mitochondrial proteins, organism clasification, bioinformatics

BARTOŇ, V. Numerické reprezentace proteinových sekvencí pro klasifikaci. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2016. 59 s. Vedoucí práce Ing. Denisa Maděránková, Ph.D.

Poděkování

Rád bych zde věnoval poděkování vedoucí této práce Ing. Denise Maděránkové, Ph.D za odborné vedení, trpělivost a podnětné rady ke zpracování této práce. Dále bych rád poděkoval svým rodičům za podporu, kterou mi ve studiu poskytují.

Prohlášení

Prohlašuji, že svou bakalářskou práci na téma Numerické reprezentace proteinových sekvencí pro klasifikaci jsem vypracoval samostatně pod vedením vedoucí bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne 24. května 2016 .....

Vojtěch Bartoň

Obsah

1. Úvod

Bioinformatika je vědní obor zabývající se metodami získávání, analýzy, vizualizací a zpracování rozsáhlých biologických dat, zejména pak z oboru molekulární biologie. Rozsáhlé soubory dat získané sekvenací jednotlivých genů, proteinů nebo celých genomů a proteomů není možné zpracovávat bez počítačové podpory [1]. Zpracování dat zapsaných písmennými kódy se využívá při zpracování genomických dat. Aplikace stejných postupů na data proteomická se již nejeví jako vhodná z důvodu nutnosti značného rozšíření používaná písmenné abecedy a tím také prudkému nárůstu všech počítaných možností a výpočetního času. Nejjednodušším řešením se tedy jeví převod textových dat na data numerická, která lze dále zpracovávat použitím matematických modelů, či signálovým zpracováním.

Vyvstává však otázka, jak zvolit vhodnou numerickou reprezentaci, respektive, které parametry daného proteinu jsou natolik významné, aby se promítly do výsledné reprezentace, a které lze naopak zanedbat. Jako výhodné se jeví pracovat s měřitelnými fyzikálněchemickými vlastnostmi jednotlivých aminokyselin, potažmo celých proteinů. Při volbě vhodné reprezentace je nutné také přihlédnout ke způsobu dalšího zpracování.

Práce je rozčleněna do dvou částí. V první části (kapitoly 2 a 3) představím různé způsoby numerických reprezentací proteinů. Pro přehlednost jsem je roztřídil do kategorií dle hlavního popisného parametru. V druhé části práce (kapitola 4) provádím testování vybraných reprezentací na reálných datech souboru mitochondriálně kódovaných proteinů obratlovců.

Veškeré programové kódy použité v práci jsou součástí přílohy elektronické verze této práce.

2. Proteomická data

Pro analýzu a klasifikaci proteomických dat jsme schopni tato data získávat přímým sekvenováním proteinů. Tato metoda je však finančně a technicky stále příliš náročná a pro účely taxonomických studií tedy nevhodná. Význam může mít snad jen při sekvenování koncové části, kde získáme krátké oligonukleotidy, které nám mohou pomoci při hledání genu v genomové knihovně. Naproti tomu jsme poměrně snadno schopni získat sekvenací data genomická, tedy posloupnost nukleotidů, která následně počítačově přeložíme na data proteomická (viz. Tabulka 2.1).

Tabulka 2.1: Genetický kód a kódování aminokyselin

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

Genetický kód organismů je degenerovaný, tedy jedna aminokyselina bývá často vyjádřena několika různými triplety. Pro klasifikaci organismů bývá často vhodnější pracovat s proteomickými daty z důvodu menšího počtu substitucí v kódu. Jelikož záměna aminokyselin naruší funkci celého proteinu, bývá takováto substituce reparačními mechanismy buňky opravena, zatímco substituce v genomických datech nemusí nutně vést ke změně funkce proteinu, protože pořadí aminokyselin bude zachováno. Sekvence proteinů v organismu bývají více evolučně konzervované. [2]

S rozmachem bioinformatiky se stalo standardní praxí zveřejňování osekvenovaných dat ve veřejně dostupných internetových databázích. V praxi již tedy není nutné každý organismus, gen, či protein zvlášť sekvenovat pro každou práci. Osekvenovaná data, často i s popisem a dalšími údaji, lze stáhnout z databáze a dále s nimi pracovat. Pro zápis dat do digitálních souborů byla vyvinuta řada formátů. Mezi bezpochyby nejrozšířenější patří GenBank a Fasta, umožňující zápis sekvence v textovém souboru, který není čitelný výhradně strojově (human-readable). Pro vyjádření sekvencí využíváme standardní IUPAC kódy. Sekvence bývají často variabilní, proto bývá vhodnější vyjádřit konsensus sekvence pomocí zápisu i nejednoznačných pozic (viz. Tabulka 2.2). [3]

Tabulka 2.2: IUPAC kódy pro zápis sekvencí.

DNA a RNA		**Protein**		
Kód	**Báze**	**Kód**	**Aminokyselina**	**Zkratka**
A	Adenin	A	Alanin	*Ala*
C	Cytosin	C	Cystein	*Cys*
G	Guanin	D	Asparát	*Asp*
T	Thymin	E	Glutamát	*Glu*
U	Uracil	F	Fenylalanin	*Phe*
R	A, G	G	Glycin	*Gly*
Y	C, T	H	Histidin	*His*
S	G, C	I	Isoleucin	*Ile*
W	A, T	K	Lysin	*Lys*
K	G, T	L	Leucin	*Leu*
M	A, C	M	Methionin	*Met*
B	C, G, T	N	Asparagin	*Asn*
D	A, G, T	P	Prolin	*Pro*
H	A, C, T	Q	Glutamin	*Gln*
V	A, C, G	R	Arginin	*Arg*
N	cokoliv	S	Serin	*Ser*
-	mezera	T	Threonin	*Thr*
		V	Valin	*Val*
		W	Tryptofan	*Trp*
		Y	Tyrosin	*Tyr*
		X	Cokoliv	*Xxx*
		B	Asparát, Asparagin	*Asx*
		Z	Glutamát, Glutamin	*Glx*

3. Numerické reprezentace

V další části textu této práce se budu zaobírat jednotlivými možnými metodami numerického vyjádření proteinů. Jednotlivé metody jsem rozdělil do kapitol, dle hlavního využitého popisného parametru.

Častým způsobem hodnocení proteinů je vizuální kontrola. Takováto kontrola je možná zejména při zobrazeních jedno až třívektorových. Při 1D reprezentaci můžeme na jednotlivé hodnoty pohlížet jako na amplitudu signálu a vynést ji do grafu s ekvidistantní osou x, vyjádřující pořadí zobrazeného prvku. Při grafickém znázornění můžeme signál vynášet kumulativně, tedy hodnota následujícího prvku se přičte k hodnotě prvku předcházejícího. Nebo budeme postupně spojovat jednotlivé body pevně dané v prostoru a tím dáme vzniknout zigzag křivce, kterou dále hodnotíme.

3.1 Reprezentace přirozenými čísly

Pro práci se soubory lze využít nejjednodušší numerickou reprezentaci. Nahrazení jednotlivých aminokyselin čísly 0-20, která vyjadřují jejich pořadí v tabulce IUPAC kódů (viz. Tabulka 2). Zavedením této reprezentace nedokážeme vyjádřit biochemické vlastnosti daného proteinu, je tedy nutné ji dále zpracovávat jinými metodami. Význam má pouze pokud je třeba vyjádřit sekvenci nikoliv posloupností písmen, ale zápisem ve vektoru čísel. [4]

Jiný přístup lze zavést pomocí mapování celých tripletů. Tím dojde k zachování informací o genomickém původu vybrané aminokyseliny. Celkem máme k dispozici šedesátčtyři různých kodónů, tedy čísla 0 až 63. Přidělování čísel vychází ze standardní tabulky kódování aminokyselin (viz. Tabulka 2.1), kde kodonům v první buňce (UUX) přidělíme postupně čísla 0-3 a poté se přesuneme do buňky napravo a přidělíme čísla 4-7 a tak dále (viz. Tabulka 3.1). [5]

Tabulka 3.1: Vyjádření kodónů přirozenými čísly.

Čísla kodónů	Číslo aminokyseliny	Název	Kód
10, 11, 14	0	Stop	stop
0, 1	1	Fenylalanin	F
2, 3, 16, 17, 18, 19	2	Leucin	L
4, 5, 6, 7, 44, 45	3	Serin	S
8, 9	4	Tyrosin	Y
12, 13	5	Cystein	C
15	6	Tryptofan	W
20, 21, 22, 23	7	Prolin	P
24, 25	8	Histidin	H
26, 27	9	Glutamin	Q
28, 29, 30, 31, 46, 47	10	Arginin	R
32, 33, 34	11	Isoleucin	I
35	12	Methionin	M
36, 37, 38, 39	13	Threonin	T
40, 41	14	Asparagin	N
42, 43	15	Lysin	K
48, 49, 50, 51	16	Valin	V
52, 53, 54, 55	17	Alanin	A
56, 57	18	Asparát	D
58, 59	19	Glutamát	E
60, 61, 62, 63	20	Glycin	G

3.2 Reprezentace EIIP hodnotami

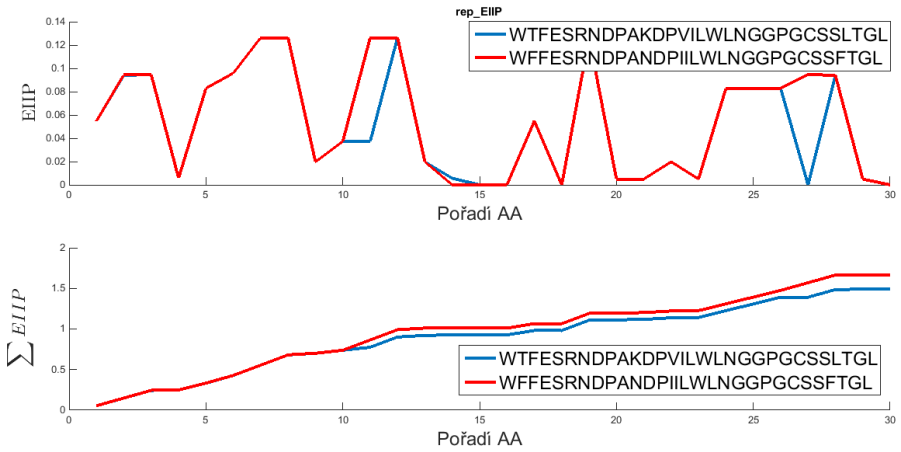
Proteinovou sekvenci lze transformovat do jednovektorové reprezentace jednoduchým nahrazením jednotlivých aminokyselin hodnotami jejich electron-ion interaction potential (EIIP). (viz. Tabulka 3.2) (viz. Obrázek 3.1) [6]

Tabulka 3.2: EIIP hodnoty aminokyselin.

aminokyselina	**A**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**
EIIP [Ry]	0,0373	0,0829	0,1263	0,0058	0,0946	0,0050	0,0242	0,0000	0,0371	0,0000
aminokyselina	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
EIIP [Ry]	0,0823	0,1263	0,0198	0,0761	0,0959	0,0829	0,0941	0,0057	0,0548	0,0516

3.3 Reprezentace založené na tetrahedronu

Tyto numerické reprezentace jsou pouze upravenou verzí již známé reprezentace pro nukleotidy. Pro jejich použití je třeba znát nukleotidovou sekvenci zpracovávaného proteinu. Výsledný vektor je značně konzervovaný a tedy je možné přecházet mezi původní a numerickou reprezentací a naopak.



Obrázek 3.1: Grafické znázornění reprezentace EIIP hodnotami.

Existují tři hlavní klasifikační příznaky pro nukleotidy, na základě kterých je lze rozdělit do tří tříd. První třída vyjadřuje molekulární strukturu, tedy zda jde o nukleotidy na bázi purinu, či pirimidinu. Druhá třída vyjadřuje vazebnou sílu, tedy počet vazeb mezi nukleotidy, a třetí třída vyjadřuje obsah amino nebo keto skupiny. Umístíme-li pomyslný tetrahedron do středu 3D souřadného kartézského systému, pak každý vrchol tetrahedronu reprezentuje jeden nukleotid a každá hrana odpovídá jedné klasifikační třídě. Každý nukleotid pak může být vyjádřen vektorem vycházejícím z bodu (0; 0; 0).

$$\begin{aligned} \vec{a} &= (\vec{i}, \vec{j}, \vec{k}) \\ \vec{c} &= (-\vec{i}, \vec{j}, -\vec{k}) \\ \vec{g} &= (-\vec{i}, -\vec{j}, \vec{k}) \\ \vec{t} &= (\vec{i}, -\vec{j}, -\vec{k}) \end{aligned} \tag{3.1}$$

Pokud toto vyjádření rozšíříme na aminokyseliny, je třeba procházet sekvenci nukleotidů po jednotlivých tripletech. Každý jednotlivý nukleotid převedeme na vektor dle výše uvedeného postupu. Výsledný vektor vyjadřující aminokyselinu získáme sečtením váhovaných vektorů jednotlivých nukleotidů. Váhy volíme jako čísla o základu 2, tedy pro poslední nukleotid nastavíme váhu na 2^0, jelikož tento nukleotid není příliš signifikantní pro určení výsledné aminokyseliny. Prostřednímu, významnějšímu vektoru nukleotidu, přidělíme váhu 2^1 a prvnímu nukleotidu v tripletu váhu 2^2, jelikož má pro určení aminokyseliny největší význam.

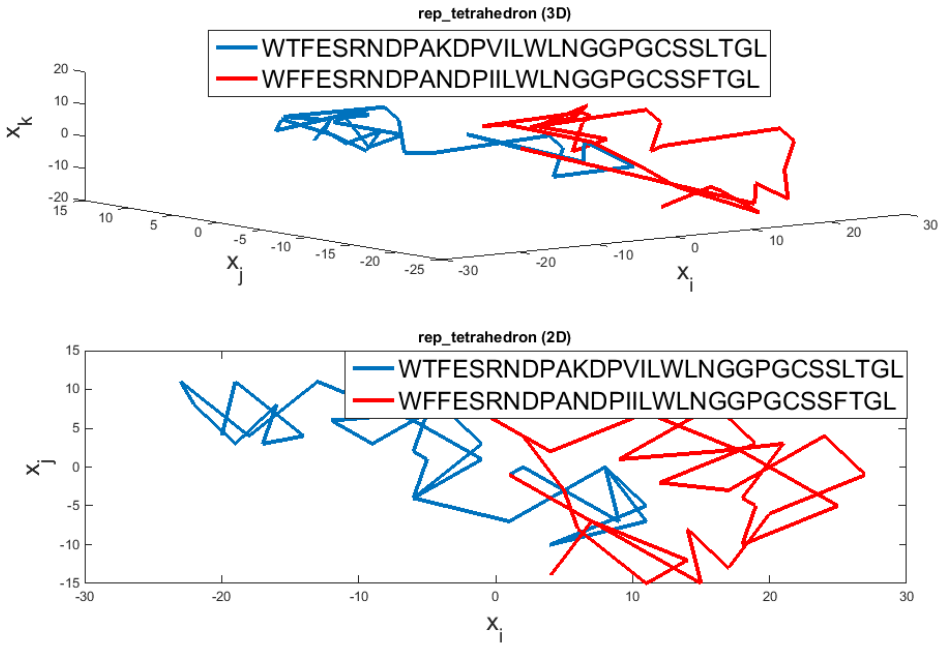
$$\vec{x} = 2^2\vec{b}_2 + 2^1\vec{b}_1 + 2^0\vec{b}_0, \quad b_i \in \{\vec{a}, \vec{b}, \vec{c}, \vec{d}\}, \quad i = \{0, 1, 2\} \tag{3.2}$$

Promítnutím nukleotidového tetrahedronu do komplexní roviny lze zredukovat rozměr reprezentujícího vektoru na 2D. Jako reálnou osu lze zvolit vektor ve směru osy x a jako imaginární osu směr osy z v 3D souřadném systému (viz. Obrázek 3.2).

$$\begin{aligned} a &= 1+j \\ c &= -1-j \\ g &= -1+j \\ t &= 1-j \end{aligned} \tag{3.3}$$

Výsledný vektor pro jednotlivé aminokyseliny získáme obdobně jako v předchozí reprezentaci, tedy váhovaným součtem vektorů jednotlivých nukleotidů. Váhy zvolíme stejně jako v předchozím případě, tedy mocniny čísla 2. [5]

$$x = 2^2 b_2 + 2^1 b_1 + 2^0 b_0, \quad b_i \in \{a, c, g, t\}, \quad i = \{0, 1, 2\} \tag{3.4}$$



Obrázek 3.2: Grafické znázornění reprezentace pomocí tetrahedronu.

3.4 Reprezentace hodnotami izoelektrického bodu

Další z možností vyjádření proteinu ve formě signálu je reprezentace jednotlivých aminokyselin jejich hodnotami izoelektrického bodu při teplotě 25 °C (viz. Tabulka 3.5). Izoelektrický

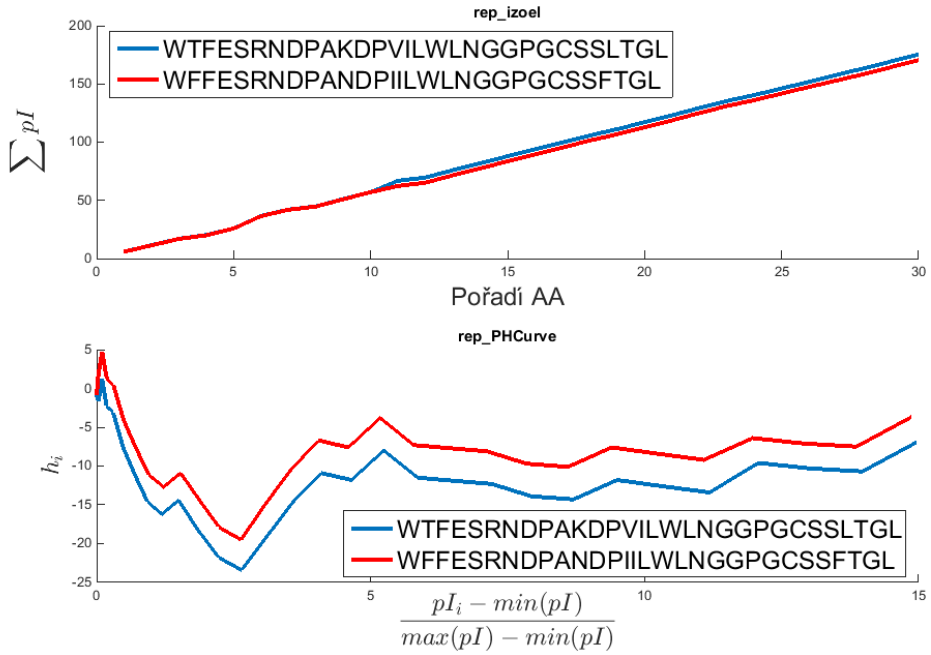
bod je hodnota pH rozpouštědla, v němž má aminokyselina nulový volný elektrický náboj. Jde o jeden z významných biochemických deskriptorů jednotlivých aminokyselin a je pro každou aminokyselinu unikátní.

Jedna z možných reprezentací je nahrazení jednotlivých aminokyselin hodnotami jejich izoelektrického bodu. Dostaneme konečně dlouhý diskrétní signál (viz. Obrázek 3.3 nahoře), který lze dále zpracovávat známými metodami signálového zpracování jako například spektrální analýzou. [7]

Druhá možná reprezentace využívající k charakterizaci proteinu hodnotu isoelektrického bodu a hydrofobicitu aminokyseliny je P-H křivka. [8] Jde o reprezentaci, kde pro každou aminokyselinu určíme její hodnoty v 2D kartézském prostoru dle vzorce 3.5 a 3.6. Jejich následným spojením vznikne křivka, která může sloužit pro grafické porovnávání s dalšími sekvencemi (viz. Obrázek 3.3 dole).

$$x_i = \frac{pI_i - min(pI)}{max(pI) - min(pI)} \tag{3.5}$$

$$y_i = h_i \tag{3.6}$$



Obrázek 3.3: Grafické znázornění reprezentace hodnotami izoelektrického bodu.

3.5 Reprezentace založené na hydrofilitě aminokyselin

Pro numerickou reprezentaci aminokyselin se často využívá detailního hydrofobně-hydrofilně polárního modelu (detailní HP model). Na základě polarity reziduí lze rozdělit aminokyseliny do čtyř kategorií: nepolární, negativně polární, nenabitě polární a pozitivně polární (viz. Tabulka 3.3).

Tabulka 3.3: Kategorizace aminokyselin dle detailního HP modelu.

Kategorie	Aminokyseliny
nepolární (np)	A, V, L, I, P, F, W, M
negativně polární (nep)	D, E
nenabitě polární (up)	G, S, T, C, Y, N, Q
pozitivně polární (pp)	K, R, H

Pro každou aminokyselinu v sekvenci je určená kategorie reprezentována vektorem B:

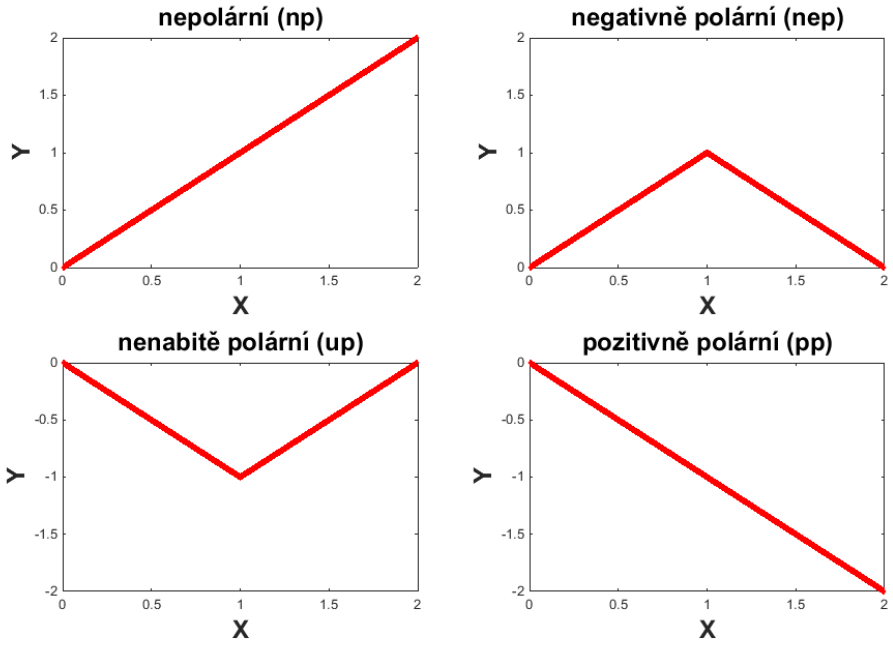
$$\begin{aligned} B_{np} &\Rightarrow (1,1),(1,1) \\ B_{nep} &\Rightarrow (1,1),(1,-1) \\ B_{up} &\Rightarrow (1,-1),(1,1) \\ B_{pp} &\Rightarrow (1,-1),(1,-1) \end{aligned} \tag{3.7}$$

Celou sekvenci převedeme na signál o dvojnásobné délce sekvence. Jednotlivé vzorky signálu získáme postupným skládáním vektorů za sebe, kde koncový bod jednoho vektoru bude výchozím bodem vektoru následujícího (viz. Obrázek 3.4). Jestliže výchozím bodem bude bod (0,0), pak jednotlivé vzorky signálu obdržíme dle vzorce:

$$\begin{aligned} pro\ i &= 1,2,3,...,n \\ S_{2i-1} &= \begin{Bmatrix} S_{2i-2}+1, & pro\ B_{np}\ a\ B_{nep} \\ bS_{2i-2}-1, & pro\ B_{up}\ a\ B_{pp} \end{Bmatrix} \\ S_{2i} &= \begin{Bmatrix} S_{2i-1}+1, & pro\ B_{np}\ a\ B_{up} \\ bS_{2i-1}-1, & pro\ B_{nep}\ a\ B_{pp} \end{Bmatrix} \end{aligned} \tag{3.8}$$

Takto získaný signál (viz. Obrázek 3.5 nahoře) můžeme dále zpracovávat postupy signálového zpracování. [9]

Další reprezentace nepoužívá ke kategorizaci aminokyselin detailní HP model, ale model jiný. Dle vlastností jednotlivých molekul aminokyselin je můžeme zjednodušeně roztřídit dle hydrofility do čtyř kategorií: silně hydrofobní, slabě hydrofobní, silně hydrofilní a slabě hydrofilní (viz. Tabulka 3.4).



Obrázek 3.4: Grafické znázornění vektorů kategorií HP modelu.

Tabulka 3.4: Zařazení aminokyselin do kategorií.

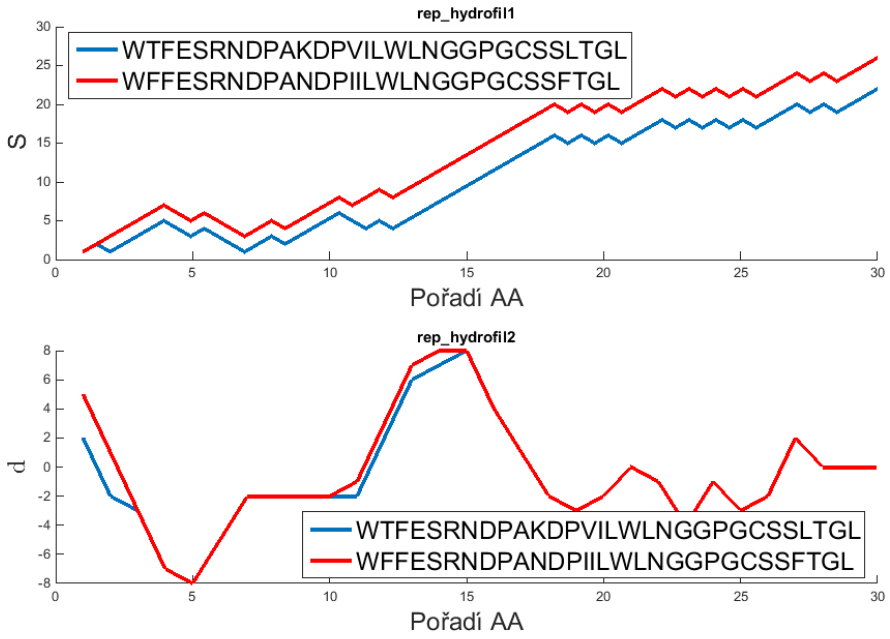
Kategorie	Aminokyseliny
silně hydrofobní (SH)	F, L, I, Y, W
slabě hydrofobní (WH)	M, V, A, P, C
slabě hydrofilní (WP)	T, H, Q, E, G
silně hydrofilní (SP)	S, N, K, D, R

Získanou sekvenci aminokyselin transformujeme do vektoru d dle vzorce:

$$d_i \quad = \quad \left\{ \begin{matrix} 2 & s_i \in SH \\ 1 & s_i \in WH \\ -1 & s_i \in WP \\ -2 & s_i \in SP \end{matrix} \right\} \tag{3.9}$$

Výslednou sekvenci S získáme ve formě ekvidistantního vektoru, kde amplituda ve směru osy y bude spočítána jako průměr čtyř po sobě jdoucích hodnot. Vzhledem k omezení výpočtu do okna o šířce čtyř aminokyselin, pohybujícího se po celé délce sekvence, bude výsledný vektor kratší o tři prvky. (viz. Obrázek 3.5 dole) [10]

$$S \quad = \quad \left\{ \begin{matrix} x_i = i; \\ y_i = \sum_{j=i}^{i+3} d_j \end{matrix} \right\}, \quad i \quad = \quad 1, 2, ..., N-3 \tag{3.10}$$



Obrázek 3.5: Grafické znázornění reprezentace založené na hydrofilitě aminokyselin.

3.6 Reprezentace založené na disociačních konstantách

Bylo popsáno několik numerických reprezentací umožňujících přiřadit každé aminokyselině její vlastní reprezentující vektor, který se dá spočítat pomocí známých fyzikálně--chemických vlastností dané aminokyseliny. Často je pak na sekvenci nahlíženo jako na 2D numerický signál.

Jedním z určujících parametrů aminokyseliny, který souvisí také s jejími výslednými vlastnostmi, je disociační konstanta. Disociační konstantu dané aminokyseliny určujeme zvlášť pro -COOH a -NH3 konec (viz. Tabulka 3.5). Tento parametr můžeme zvolit jako jeden ze základních parametrů pro tvorbu numerické reprezentace, jelikož determinuje aktivitu enzymů a dalších biochemických vlastností.

Tabulka 3.5: Disociační konstanty, izoelektrický bod a index hydrofobicity.

Aminokyselina	**Kód**	**pKa -COOH**	**pKa -NH3**	**pI**	**h**
Alanin	A	2,34	9,69	6,00	1,80
Cystein	C	1,71	10,78	5,07	2,50
Asparát	D	2,09	9,82	2,77	-3,50
Glutamát	E	2,19	9,67	3,22	-3,50
Fenylalanin	F	1,83	9,13	5,48	2,80
Glycin	G	2,34	9,60	5,97	-0,40
Histidin	H	1,82	9,17	7,59	-3,20
Isoleucin	I	2,36	9,68	6,02	4,50
Lysin	K	2,18	8,95	9,74	-3,90
Leucin	L	2,36	9,60	5,98	3,80
Methionin	M	2,28	9,21	5,74	1,90
Asparagin	N	2,02	8,80	5,41	-3,50
Prolin	P	1,99	10,60	6,30	-1,60
Glutamin	Q	2,17	9,13	5,65	-3,50
Arginin	R	2,17	9,04	10,76	-4,50
Serin	S	2,21	9,15	5,68	-0,80
Threonin	T	2,63	10,43	5,60	-0,70
Valin	V	2,32	9,62	5,96	4,20
Tryptofan	W	2,38	9,39	5,89	-0,90
Tyrosin	Y	2,20	9,11	5,66	-1,30

Reprezentace založená na hodnotách disociačních konstant byla navržena v Chemical Physics Letters [11]. Výstupem bude 2D signál, kde každá aminokyselina je reprezentována vlastním vektorem. První souřadnice vektoru odpovídá hodnotě disociační konstanty COOH-konce aminokyseliny. Druhou souřadnici vypočítáme jako rozdíl hodnoty disociační konstanty NH3 konce a aritmetického průměru hodnot disociačních konstant NH3 konce všech aminokyselin:

$$b_i = (pK_a(COOH)_i, \quad pK_a(NH_3) - \overline{pK_a(NH_3)}). \tag{3.11}$$

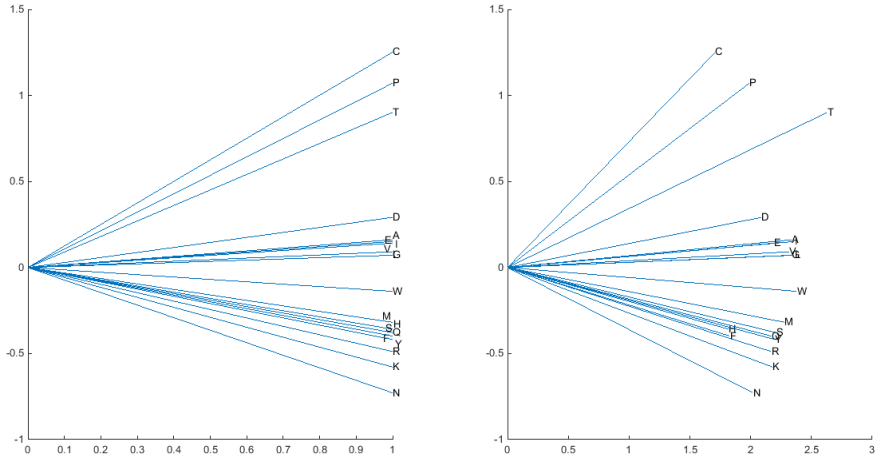
Výsledný signál vytvoříme opět skládáním vektorů za sebe, kde koncový bod jednoho je počátečním bodem vektoru následujícího (viz. Obrázek 3.7 nahoře).

Další používaná reprezentace navrhuje využít hodnoty disociačních konstant přímo jako souřadnice jednotlivých aminokyselin. Jako vhodné se prokázalo odečtení průměrné hodnoty disociační konstanty všech dvaceti aminokyselin od hodnoty uvažované aminokyseliny. Tím získáme hodnoty rozložené na obou stranách číselné osy.

$$
\begin{aligned}
x_i &= pK_a(COOH)_i - \overline{pK_a(COOH)}; \\
y_i &= pK_a(NH_3)_i - \overline{pK_a(NH_3)}.
\end{aligned}
\tag{3.12}
$$

Pokud takto převedenou posloupnost aminokyselin vyneseme do grafu a interpolujeme fraktálovou metodou, obdržíme F-křivku, vhodnou ke grafickému tvarovému porovnávání proteinových sekvencí. [12]

Při grafickém porovnávání takto vytvořených křivek proteinových sekvencí může docházet ke zkreslení vlivem nestejně velké hodnoty x-souřadnice. Ta v případě stejně dlouhých sekvencí, může zapříčinit různě dlouhou křivku, což ztíží lokální porovnání některých úseků. V některých případech je proto vhodnější nastavit x-souřadnici neměnnou hodnotu (viz. Obrázek 3.6).

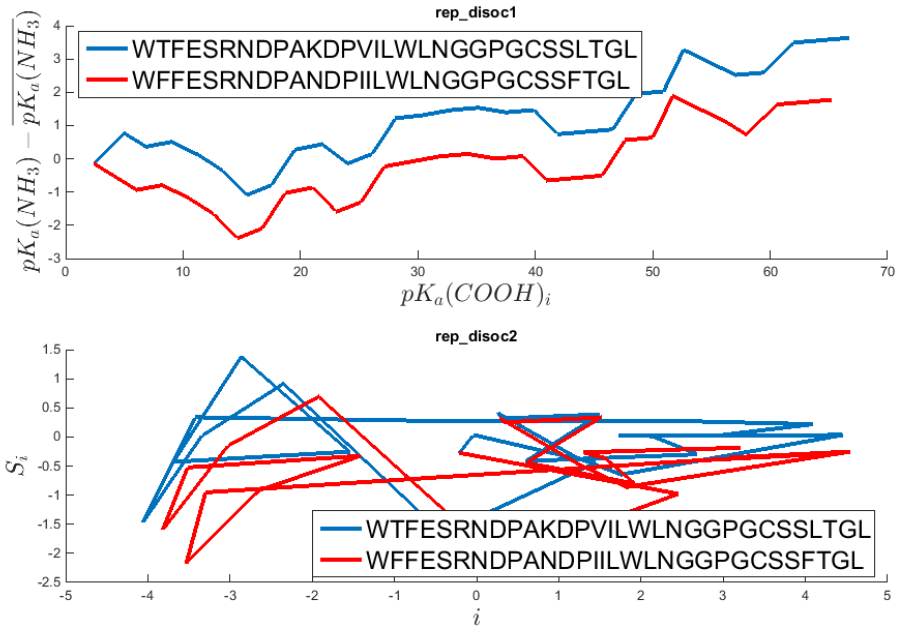


Obrázek 3.6: Ekvidistantní a proměnná x-souřadnice reprezentace.

Disociační konstanta však není jediným významný popisným parametrem aminokyselin. Do výpočtu můžeme zahrnout také další popisné parametry jako hodnotu izoelektrického bodu nebo stupeň hydrofobicity. Taková reprezentace byla navržena například v článku A graphical representation of protein based on a novel iterated function system [13] nebo Novel Numerical Characterization of Protein Sequences Based on Individual Amino Acid and Its Application [14].

V těchto článcích je každá aminokyselina opět vyjádřena svým unikátním vektorem P, kde při výpočtu jeho souřadnic zohledňujeme i další parametry. Výsledná sekvence bodů S už není prostým skládáním původních vektorů, ale v potaz je brána také aminokyselina předcházející, jelikož i ta svými vlastnostmi ovlivňuje vlastnosti okolních aminokyselin. Výsledná 2D reprezentace bude počítána podle vzorce 3.13, kde S_0 je (0; 0) (viz. Obrázek 3.7 dole).

$$
\begin{aligned}
P_{x_i} &= pK_a(COOH)_i - \overline{pK_a(COOH)} + h_i - \overline{h} \\
P_{y_i} &= pK_a(NH_3)_i - \overline{pK_a(NH_3)} + pI_i - \overline{pI} \\
S_i &= \begin{Bmatrix} x_i = \frac{3}{4}x_{i-1} + \frac{1}{2}P_{x_i} \\ y_i = \frac{3}{4}y_{i-1} + \frac{1}{2}P_{y_i} \end{Bmatrix}
\end{aligned} \tag{3.13}
$$



Obrázek 3.7: Grafické znázornění reprezentace založené na disociačních konstantách.

Při reprezentacích vhodných k vytvoření grafického výstupu se nemusíme omezovat pouze na 2D prostor. Pro práci jsem vybral reprezentaci rozšiřující vzorec 3.12 o výpočet souřadnice třetí: [15]

$$
\begin{aligned}
x_i &= pK_a(COOH)_i - \overline{pK_a(COOH)}; \\
y_i &= pK_a(NH_3)_i - \overline{pK_a(NH_3)}; \\
y_i &= h_i - \overline{h}
\end{aligned} \tag{3.14}
$$

Transformaci provedeme pro každou aminokyselinu v sekvenci zvlášť, čímž obdržíme souřadnice jednotlivých bodů v 3D prostoru. Posloupnost bodů poté upravíme dle vzorce 3.15, tak abychom obdrželi kontinuální diskrétní signál vhodný k další analýze.

$$
\begin{aligned}
X_i &= \sum_{k=1}^{i} x_k \\
Y_i &= \sum_{k=1}^{i} y_k \\
Z_i &= \sum_{k=1}^{i} z_k
\end{aligned}
\tag{3.15}
$$

3.7 Binární vyjádření založené na metodě Huffmanova stromu

Jde o dvouvektorové vyjádření založené na binárním zakódování jednotlivých aminokyselin. Vyjádření aminokyselin pomocí jedniček a nul umožňuje kompresi celkové informace s ohledem na frekvenci zastoupení jednotlivých aminokyselin v sekvenci.

Huffmanovo kódování [16] je metodou digitální komprese dat. Je to metoda s proměnnou délkou kódování, závisející na kódovaném souboru dat. Délka jednoho kódu závisí na frekvenci výskytu zvoleného symbolu v souboru. Huffmanovo kódování můžeme rozdělit na dynamické, kde není dopředu známá frekvence výskytu symbolů, a statické, kde máme informaci o frekvenci výskytu symbolů předem danou. Pro transformaci proteinové sekvence můžeme využít statickou metodu kódování, neboť kódovanou sekvenci již známe.

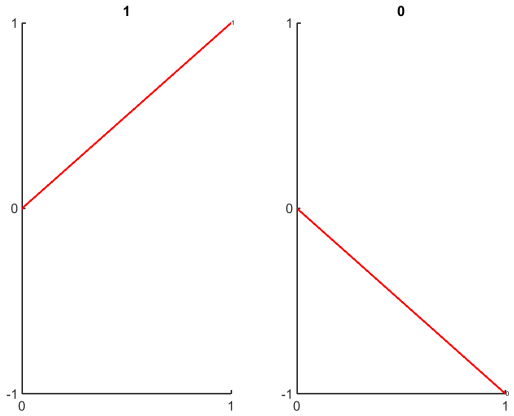
Huffmanův strom je striktně binárním stromem. Striktně binární stromy jsou speciálním typem binárních stromů, ve kterých mají všechny uzly, až na koncové, právě dva potomky, pravého a levého.

Pro vytvoření Huffmanova stromu z našich dat musíme nejprve získat veškeré sekvence, které budeme porovnávat. Následně zjistíme početní zastoupení jednotlivých aminokyselin v celém souboru dat. Najdeme prvky s nejnižší frekvencí výskytu a spojíme je do uzlu. Následně přepočítáme tabulku frekvencí tak, že prvky v uzlu spojíme a jejich frekvence sečteme. Poté opět hledáme prvky s nejnižší frekvencí a spojíme je do uzlu. Postup opakujeme dokud nebudou všechny prvky spojené do jednoho kořenového uzlu.

Nyní máme binární strom s kořenovým uzlem vyjadřujícím celkový počet aminokyselin. Dále postupujeme od kořene stromu a jednotlivým větvím stromu přiřazujeme hodnoty nula a jedna tak, aby každá větev vystupující z uzlu doleva měla hodnotu nula a každá větev směřující doprava hodnotu jedna. Strom následně čteme od kořene ke koncovým uzlům vyjadřujícím jednotlivé aminokyseliny. Kód každé aminokyseliny je vytvořen po-

stupným skládáním hodnot větví ukazujících cestu ke kódované aminokyselině. Každá aminokyselina tak obdrží svou binární hodnotu a do binárního kódu je převedena celá sekvence prostým nahrazením symbolů jejich kódy.

Pro grafické znázornění a zhodnocení zakódovaných sekvencí můžeme vyjádřit obě hodnoty jejich vektory umístěnými v jednotkovém souřadném systému; do prvního kvadrantu pro hodnotu jedna a do čtvrtého kvadrantu pro hodnotu nula (viz. Obrázek 3.8). Výsledný signál získáme spojením jednotlivých vektorů za sebou. [17]



Obrázek 3.8: Vektorové vyjádření binárních hodnot 1 a 0.

3.8 Reprezentace založené na teorii chaosu

V této kapitole si představíme několik základních přístupů k numerickým reprezentacím proteinových sekvencí využívajících teorii chaosu. V odborné literatuře můžeme následující reprezentace najít sdružené pod pojmem chaos game representation (CGR).

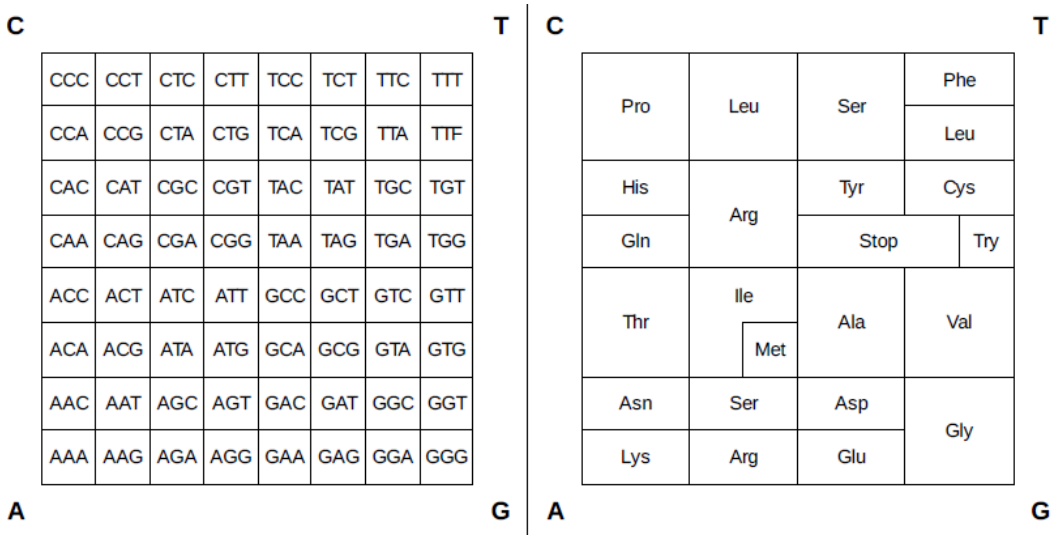
Jeden z prvotních algoritmů využíval „magického čtverce“. [18] Jde o čtverec, pro numerické vyjádření umístěný do souřadného systému s těžištěm v počátku systému a délkou hrany o dvojnásobku jednotkové délky. Původně byl navržen pro reprezentace nukleotidových sekvencí, později byl algoritmus upraven i pro proteinové sekvence. Každý vrchol čtverce představuje jinou bázi (A, C, G, T). Báze A je umístěna na vrchol ve třetím kvadrantu a následně ve směru hodinových ručiček umístíme postupně báze C, G a T.

Do oblasti vymezené naším čtvercem vykreslujeme body, tedy jednotlivé nukleotidy. Vycházíme ze středu systému (bod [0,0]), další bod umístíme do poloviny vzdálenosti k příslušnému vrcholu čtverce, dle vykreslovaného nukleotidu. Následně si zvolíme tento vykreslený bod jako výchozí a postup opakujeme v celé délce sekvence.

Výše uvedený algoritmus lze uplatnit i na reprezentaci celých kodonů. Nejjednodušším způsobem je prosté vykreslování pouze každého třetího bodu. Pro zlepšení efektivity algoritmu jej však musíme více poupravit. O tom, kterou aminokyselinu triplet reprezentuje nerozhodují všechny nukleotidy stejnou vahou, určující je jejich pořadí. Proto i pořadí nukleotidů v algoritmu přiřadíme rozdílnou váhu.

První krok zůstane neměnný a první nukleotid nám určí posun do poloviny vzdálenosti k příslušnému vrcholu čtverce. Tento posun nás přenesl do pomyslného středu čtverce menšího, ale stejně koncipovaného. Opět se posuneme o polovinu vzdálenosti k vrcholu tentokráte však menšího pomyslného čtverce. Tento posun nás opět přesunul po polovině diagonály do pomyslného středu nejmenšího čtverce. Dle typu třetí báze již vykreslíme bod v polovině vzdálenosti k příslušnému vrcholu nejmenšího čtverce. U dalšího zobrazovaného kodonu vycházíme z předchozího bodu, uvažovat však budeme opět původní největší čtverec a algoritmus obdobně opakujeme. [18]

Uvedený postup však vyžaduje, aby byla dopředu známá nukleotidová sekvence daného proteinu. Z modifikovaného algoritmu můžeme sestrojit mapu 8x8 všech 64 kodonů, kde bod reprezentující každý jednotlivý kodon náleží právě jedné buňce. Pohledem na mapu zjistíme, že kodony jsou organizovány tak, že v prvním kvadrantu leží všechny začínající bazí T, ve druhém všechny začínající bazí C, atd... (viz. Obrázek 3.9).



Obrázek 3.9: Rozložení aminokyselin v "magickém čtverci".

Souřadnice pro vyjádření jednotlivých aminokyselin určíme jako umístění těžiště oblastí se stejnou aminokyselinou (viz. Tabulka 3.6).

Pro aminokyseliny můžeme magický čtverec rozšířit na dvacetistranný polygon, po jehož obvodu rovnoměrně rozmístíme jednotlivé aminokyseliny. Začínáme alaninem umístě-

ným na polopřímku oddělující první a čtvrtý kvadrant a postupujeme dále proti směru hodinových ručiček dle abecedy, konče valinem. Z vyobrazeného polygonu umístěného do jednotkového souřadného systému obdržíme souřadnice pro jednotlivé aminokyseliny (viz. Tabulka 3.6).

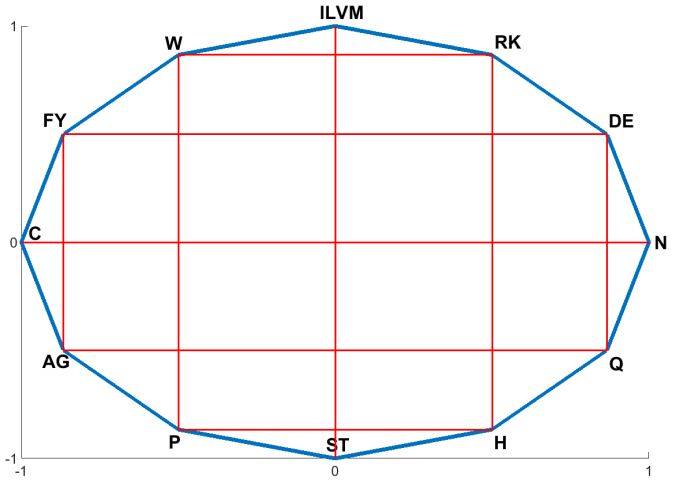
Tabulka 3.6: Souřadnice aminokyselin odvozené ze čtverce (X, Y) a z polygonu (x, y, radiany). [19]

Aminokyselina	**X**	**Y**	**x**	**y**	**Radiány**
A	+1/4	-1/4	1	0	0
R	-1/4	-1/8	0,9511	0,3090	0,3142
N	-3/4	-5/8	0,8090	0,5878	0,6283
D	+1/4	-5/8	0,5878	0,8090	0,9425
C	+3/4	+3/8	0,3090	0,9511	1,2566
Q	-3/4	+1/8	0	1	1,5708
E	+1/4	-7/8	-0,3090	0,9511	1,8850
G	+3/4	-3/4	-0,5878	0,8090	2,1991
H	-3/4	+3/8	-0,8090	0,5878	2,5133
I	-7/24	-5/24	-0,9511	0,3090	2,8274
L	+1/6	+17/24	-1	0	3,1416
K	-3/4	-7/8	-0,9511	-0,3090	3,4557
M	-1/8	-3/8	-0,8090	-0,5878	3,7699
F	+3/4	+7/8	-0,5878	-0,8090	4,0841
P	-3/4	+3/4	-0,3090	-0,9511	4,3982
S	+1/12	+7/24	0	-1	4,7124
T	-3/4	-1/4	0,3090	-0,9511	5,0265
W	+7/8	+1/8	0,5878	-0,8090	5,3407
Y	+1/4	+3/8	0,8090	-0,5878	5,6549
V	+3/4	-1/4	0,9511	-0,3090	5,9690

Pro hodnocení proteinových struktur můžeme využít jak reprezentace vyvozené ze zjištěných souřadnic, tak také analýzu vykreslených křivek a bodů přímo do čtverce, či polygonu.

Často bývá využíváno také dvanáctivrcholového polygonu aminokyselinami sdruženými tak, že jednomu vrcholu polygonu může náležet současně více aminokyselin. Ty jsou sdruženy nebo umístěny tak, aby ve společném vrcholu ležely aminokyseliny u kterých dochází často k vzájemné substituci (viz. Obrázek 3.10).

Využijeme výše zmíněného algoritmu umístění bodu do poloviny vzdálenosti k příslušnému vrcholu. Po vykreslení celé sekvence rozdělíme polygon do 24 oblastí a spočítáme procentuální zastoupení bodů v jednotlivých oblastech. Tyto hodnoty bývají velmi podobné u příbuzných proteinů. [20]

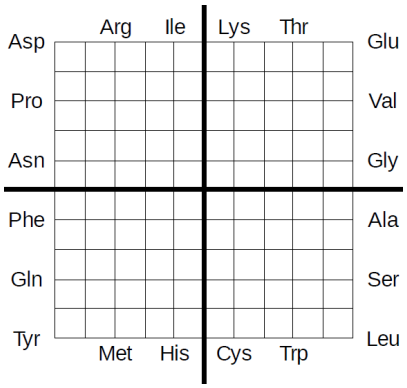


Obrázek 3.10: Dvanáctivrcholový polygon s vyznačenými oblastmi.

3.9 Reprezentace založené na mřížce

Tyto jednoduché reprezentace byly vytvořeny především pro grafickou analýzu sekvencí. Aminokyselinám přiřazují souřadnice o přirozených hodnotách v 2D souřadném prostoru.

Začneme s čtvercem 10x10 umístěném do středu souřadného systému. Po jeho obvodu rovnoměrně rozmístíme body, kterým následně přiřadíme jednotlivé aminokyseliny. Aminokyseliny budeme přiřazovat podle jejich relativního výskytu v sekvencích. Začneme tedy leucinem pravém spodním rohu a postupujeme proti směru hodinových ručiček (viz. Obrázek 3.11). Souřadnice pro reprezentaci jednotlivých aminokyselin určíme odečtením souřadnic ze čtverce.



Obrázek 3.11: Rozmístění aminokyselin v mřížce.

Toto řazení umožní částečně zamezit křížení výsledné křivky, neboť aminokyselinám s častějším výskytem jsou přiřazeny kladné hodnoty x-souřadnice (viz. Tabulka 3.7 (X, Y)).

Křížení křivky je sice omezeno, ale není zcela vyloučeno. Aby jsme jeje zcela vyloučili, je třeba přiřadit všem aminokyselinám kladné x-souřadnice. Uvažujme opět čtverec 10x10 umístěný ve středu souřadného systému. Místa přiřazovaná aminokyselinám nyní rozmístíme pouze v pravé polovině čtverce, a souřadnice odečteme (viz. Tabulka 3.7 (x, y)).

Pro popis takto sestrojených sekvencí můžeme využít také konstrukce matice vzdáleností 20x20. Najdeme výskyt jednotlivých aminokyselin v sekvenci a spočítáme jejich průměrné souřadnice. Obdržíme sérii bodů, jejichž vzdálenost změříme, nejčastěji za vzdálenost považujeme euklidovskou. Výsledná matice vzdáleností popisuje daný protein a může být využita pro srovnávací analýzu.

Tabulka 3.7: Relativní výskyt aminokyselin a jejich souřadnice dle mřížky.

Aminokyselina	**Relativní výskyt [%]**	**(X, Y)**	**(x, y)**
L	9,36	(5, -5)	(0, 5)
S	7,38	(5, -3)	(1, 5)
A	7,34	(5, -1)	(2, 5)
G	6,89	(5, 1)	(3, 5)
V	6,48	(5, 3)	(4, 5)
E	6,22	(5, 5)	(5, 5)
T	5,85	(3, 5)	(5, 4)
K	5,81	(1, 5)	(5, 3)
I	5,76	(-1, 5)	(5, 2)
R	5,20	(-3, 5)	(5, 1)
D	5,12	(-5, 5)	(5, 0)
P	5,00	(-5, 3)	(5, -1)
N	4,57	(-5, 1)	(5, -2)
F	4,12	(-5, -1)	(5, -3)
Q	3,96	(-5, -3)	(5, -4)
Y	3,25	(-5, -5)	(5, -5)
M	2,32	(-3, -5)	(4, -5)
H	2,26	(-1, -5)	(3, -5)
C	1,76	(1, -5)	(2, -5)
W	1,34	(3, -5)	(1, -5)

4. Testování reprezentací

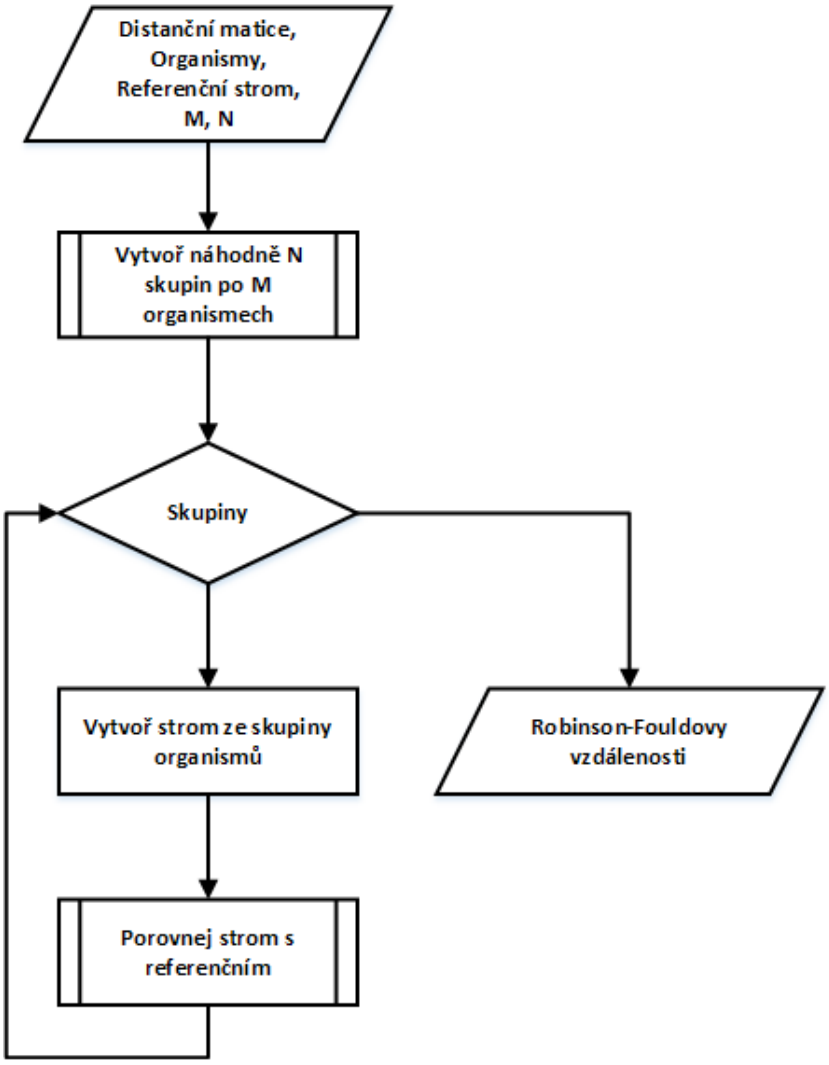
4.1 Naprogramované reprezentace

Na základě vypracované rešerše metod numerických reprezentací proteinových sekvencí bylo vybráno celkem 5 metod k realizaci a testování na reálných datech. Pro realizaci bylo zvoleno programové prostředí Matlab a programové kódy metod jsou součástí elektronické přílohy práce. Realizovány byly tyto metody:

- Reprezentace EIIP hodnotami popsaná v kapitole 3.2, v kumulované i nekumulované formě **(rep_eiip)**.
- Reprezentace hodnotami izoelektrického bodu popsaná v druhém odstavci kapitoly 3.4, v kumulované i nekumulované formě **(rep_izoel)**.
- Reprezentace hodnotami izoelektrického bodu popsaná v kapitole 3.4 vzorcem 3.5 **(rep_PHCurve)**.
- Reprezentace založená na hydrofilitě aminokyselin popsaná v kapitole 3.5 vzorcem 3.9 a 3.10 s nastavitelnou délkou okna **(rep_hydrofil2)**.
- Reprezentace založená na disociačních konstantách popsaná v kapitole 3.6 vzorcem 3.11, v kumulované i nekumulované formě **(rep_disoc1)**.

4.2 Metodika testování kvality

Numerické reprezentace pro fylogenetickou klasifikaci organismů by měly odrážet jejich taxonomické zařazení. To znamená, čím taxonomicky vzdálenější organismy, tím by měla být větší vzájemná vzdálenost numerických signálů těchto organismů. Kvalita numerické reprezentace tedy závisí na kombinaci vhodně zvolené reprezentace a na způsobu výpočtu vzdálenosti mezi nimi. Způsoby výpočtu vzdáleností se zabývá kapitola 4.3.



Obrázek 4.1: Metoda testování kvality reprezentace.

Předmětem testu kvality je vzájemná kombinace reprezentace a distanční metody. Testovaný soubor převedeme do numerické reprezentace a výsledný signál převzorkujeme tak, aby testované signály měly stejnou délku. Převzorkování zavádíme z důvodu následného výpočtu vzájemné vzdálenosti, která je většinou definována pro stejně dlouhé

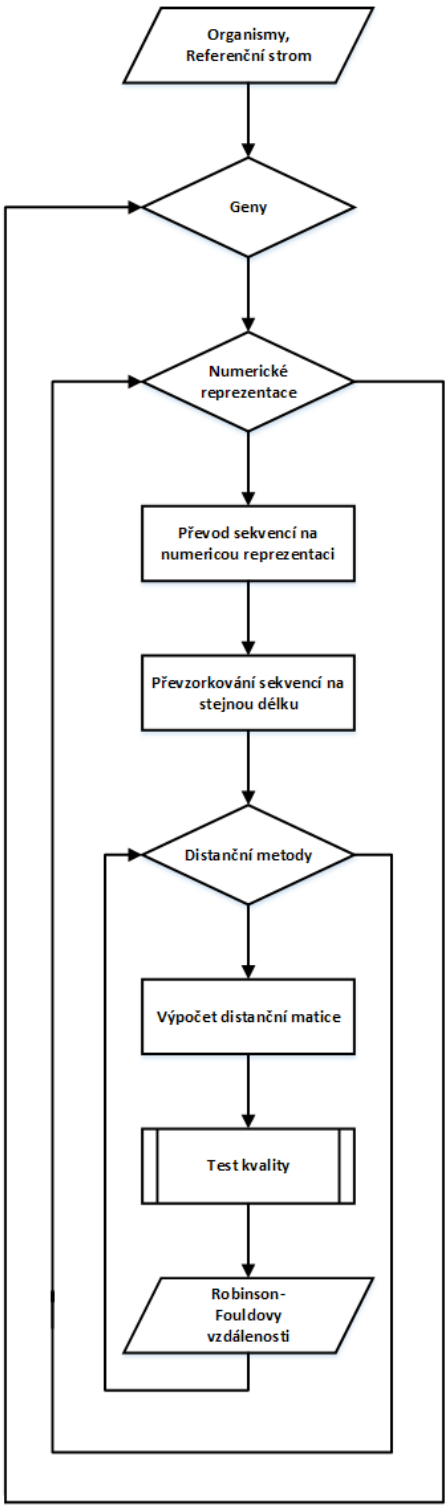
signály. Nyní vypočítáme vzájemnou vzdálenost pro každou dvojici signálů z testovaného souboru. Výsledné hodnoty vzdáleností zapisujeme do distanční matice symetrické podle diagonály. Tuto matici použijeme pro sestrojení fylogenetického stromu.

Ze souboru vybereme N organismů, pro které sestrojíme fylogenetický strom z distanční matice upravené tak, aby obsahovala pouze vybrané organismy. Tento vytvořený strom porovnáme se standardní taxonomií vybraných organismů. Pro porovnání jsme zvolili metodu Robinson-Fouldovy vzdálenosti (RF vzdálenost) [21].

Robinson-Fouldova vzdálenost je metoda porovnávání dvou dendrogramů udávající počet rozdílných shluků v obou stromech ku celkovému počtu shluků. Výsledkem je skóre v rozmezí 0 až 1. Minimální hodnoty nabývá v případě identických stromů.

Tento postup s výběrem organismů opakujeme M-krát. Výsledná skóre následně zprůměrujeme. Čím větší počet průměrovaných hodnot, tím reprezentativnější výsledek. Celé schéma testování viz. Obrázek 4.1.

Výše popsaný postup slouží pouze k otestování kombinace jedné reprezentace jednoho genu a jedné distanční metody. Pro určení nejvhodnější distanční metody pro zvolenou reprezentaci a zvolený gen je třeba testování provést pro všechny možné kombinace reprezentací a distančních metod. Ideálně pro několik genů, aby jsme je mohli srovnávat. Toto rozšíření ukazuje schéma viz. Obrázek 4.2.



Obrázek 4.2: Postup při vyhodnocování kvality reprezentací.

4.3 Distanční metody

Pro určení vzájemné vzdálenosti dvou konečných signálů používáme distanční metody. Jde o matematické vyjádření míry odlišnosti. Dále uvádím přehled v práci použitých metod výpočtu vzdálenosti [22].

Euklidova vzdálenost

$$d_{st}^2 = (x_s - x_t)(x_s - x_t)' \tag{4.1}$$

Standardizovaná euklidova vzdálenost

$$d_{st}^2 = (x_s - x_t)V^{-1}(x_s - x_t)' \tag{4.2}$$

Cityblock vzdálenost

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - x_{tj}| \tag{4.3}$$

Chebychevova vzdálenost

$$d_{st} = max_j\{|x_{sj} - x_{tj}|\} \tag{4.4}$$

Cosinova vzdálenost

$$d_{st} = 1 - \frac{x_s x_t'}{\sqrt{(x_s x_s')(x_t x_t')}} \tag{4.5}$$

Korelační vzdálenost

$$d_{st} = 1 - \frac{(x_s - \overline{x_s})(x_t - \overline{x_t})'}{\sqrt{(x_s - \overline{x_s})(x_s - \overline{x_s})'}\sqrt{(x_t - \overline{x_t})(x_t - \overline{x_t})'}}, \tag{4.6}$$
$$kde \quad \overline{x_s} = \frac{1}{n}\sum_j x_{sj} \quad a \quad \overline{x_t} = \frac{1}{n}\sum_j x_{tj}$$

Hammingova vzdálenost

$$d_{st} = (\#(x_{sj} \neq x_{tj})/n) \tag{4.7}$$

Spearmanova vzdálenost

$$\begin{gathered} d_{st} = 1 - \frac{(r_s - \overline{r_s})(r_t - \overline{r_t})'}{\sqrt{(r_s - \overline{r_s})(r_s - \overline{r_s})'}\sqrt{(r_t - \overline{r_t})(r_t - \overline{r_t})'}}, \\ kde \;\; \overline{r_s} = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2}, \\ \overline{r_t} = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2}. \end{gathered} \tag{4.8}$$

Jaccardova vzdálenost

$$d_{st} = \frac{\#[(x_{sj} \neq x_{tj}) \cap ((x_{sj} \neq 0) \cup (x_{tj} \neq 0))]}{\#[(x_{sj} \neq 0) \cup (x_{tj} \neq 0)]} \tag{4.9}$$

4.4 Zvolená množina dat

Pro testování kvality navržených reprezentací jsem zvolil soubor dat mitochondriálních genů organismů získaných z volně dostupné databáze NCBI poskytnutých vedoucím bakalářské práce. Záznamy obsahují kódující úseky pro geny ATP6, ATP8, CYTB, COX1, COX2, COX3, ND1, ND2, ND3, ND4, ND4L, ND5 a ND6. Ze souboru dat jsem vyčlenil pouze obratlovce a dále odstranil záznamy neobsahující všech třináct sledovaných genů a organismy vyskytující se v souboru vícekrát. Výsledný testovací soubor čítá 2410 organismů. V příloze elektronické verze této práce jsou dostupné soubory jednotlivých genů ve formátu fasta.

Pro celý soubor organismů jsem vytvořil standardní taxonomický strom pomocí online nástroje na stránkách NCBI [23]. Získaný strom použiji jako referenční pro porovnávání se stromy vytvořenými z numerických reprezentací. Pro použití v programovém řešení byl strom mírně upraven, ze stromu byly odstraněny duplikátní identifikátory tříd organismů a ve jménech organismů jsem nahradil mezery podtržítky pro snadnější výpočetní zpracování.

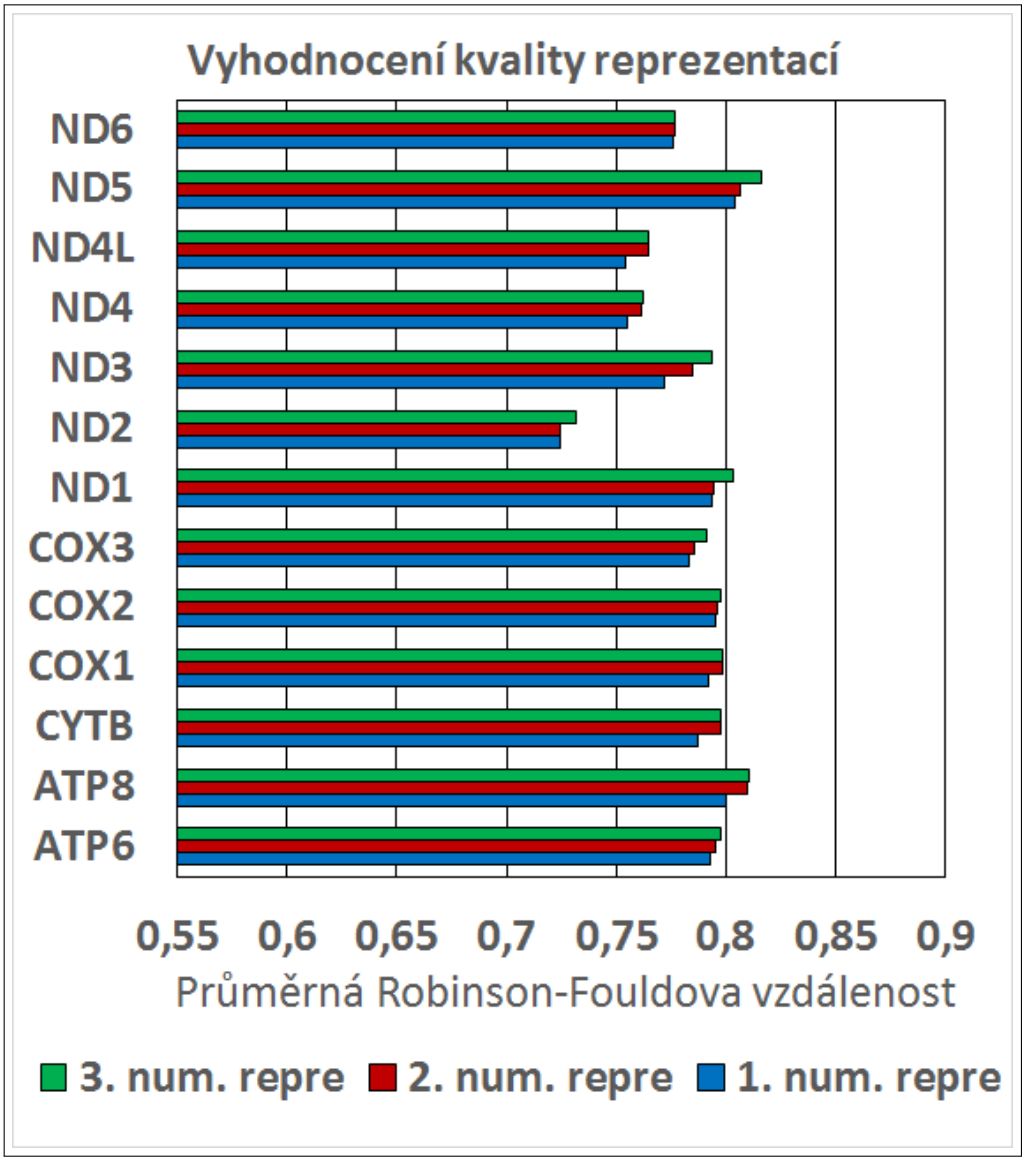
Pro konstrukci stromů ze vzdáleností numerických sekvencí jsem zvolil metodu UPGMA [24]. Volitelný parametr testu udávající velikost vytvářených stromů N jsem zvolil 20. Parametr udávající počet vytvářených stromů M jsem nastavil na 500.

Pro účely porovnání se symbolickou reprezentací jsem všechny zmíněné sekvence jednotlivých genů globálně zarovnal algoritmem ClustalOmega [25] aplikovaným v balíku msa [26] pro R [27]. Zarovnané sekvence jsou rovněž dostupné ve fasta formátu v příloze elektronické verze této práce.

4.5 Vyhodnocení kvality reprezentací

Výše popsaným postupem jsem otestoval všech třináct mitochondriálních genů celkem pro třináct reprezentací a každou reprezentaci při použití každé z devíti zvolených distančních metod. Výsledné skóre jednotlivých testů přikládám v příloze A, Tabulky A.3-15.

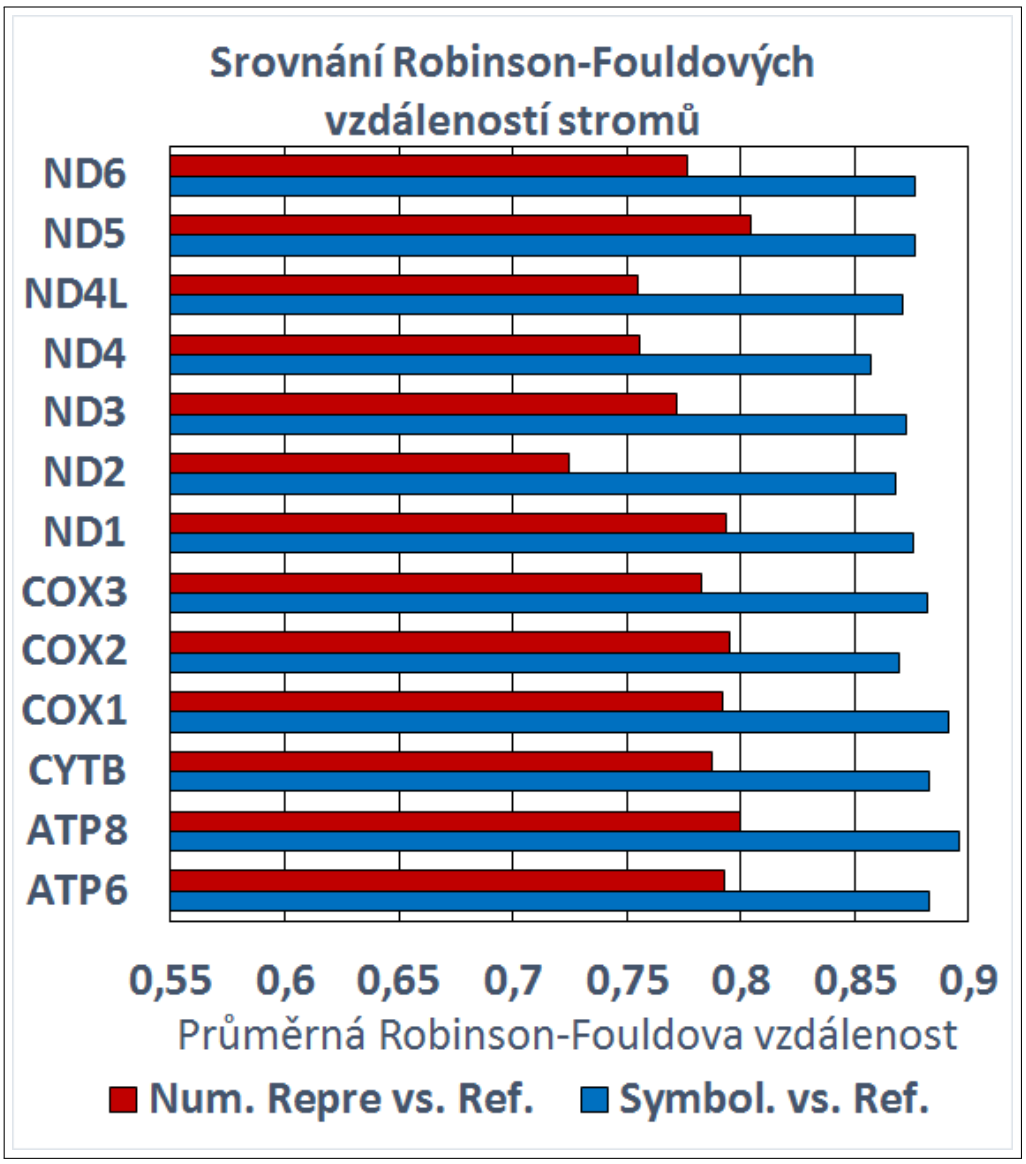
Pro vyhodnocení jsem vybral pro každý gen tři kombinace reprezentace a distanční metody o nejnižším průměrném skóre (viz. Tabulka A.2).



Obrázek 4.3: Srovnání RF vzdáleností nejvhodnějších reprezentací od standardní taxonomie.

Z grafu (viz. Obrázek 4.3) se jako nejvhodnější protein pro klasifikaci jeví protein ND2, který má nejnižší Robinson-Fouldovu vzdálenost od standardní taxonomie a to 72,46 %.

Další vhodné proteiny jsou ND4 a ND4L jejichž průměrné vzdálenosti se pohybují kolem 75 %. Z ostatních testovaných proteinů se většina pohybuje okolo 80 %. Při porovnávání se standardním taxonomickým stromem musíme brát v potaz, že strom nemusí korespondovat se vzájemnou podobností jednotlivých genů.



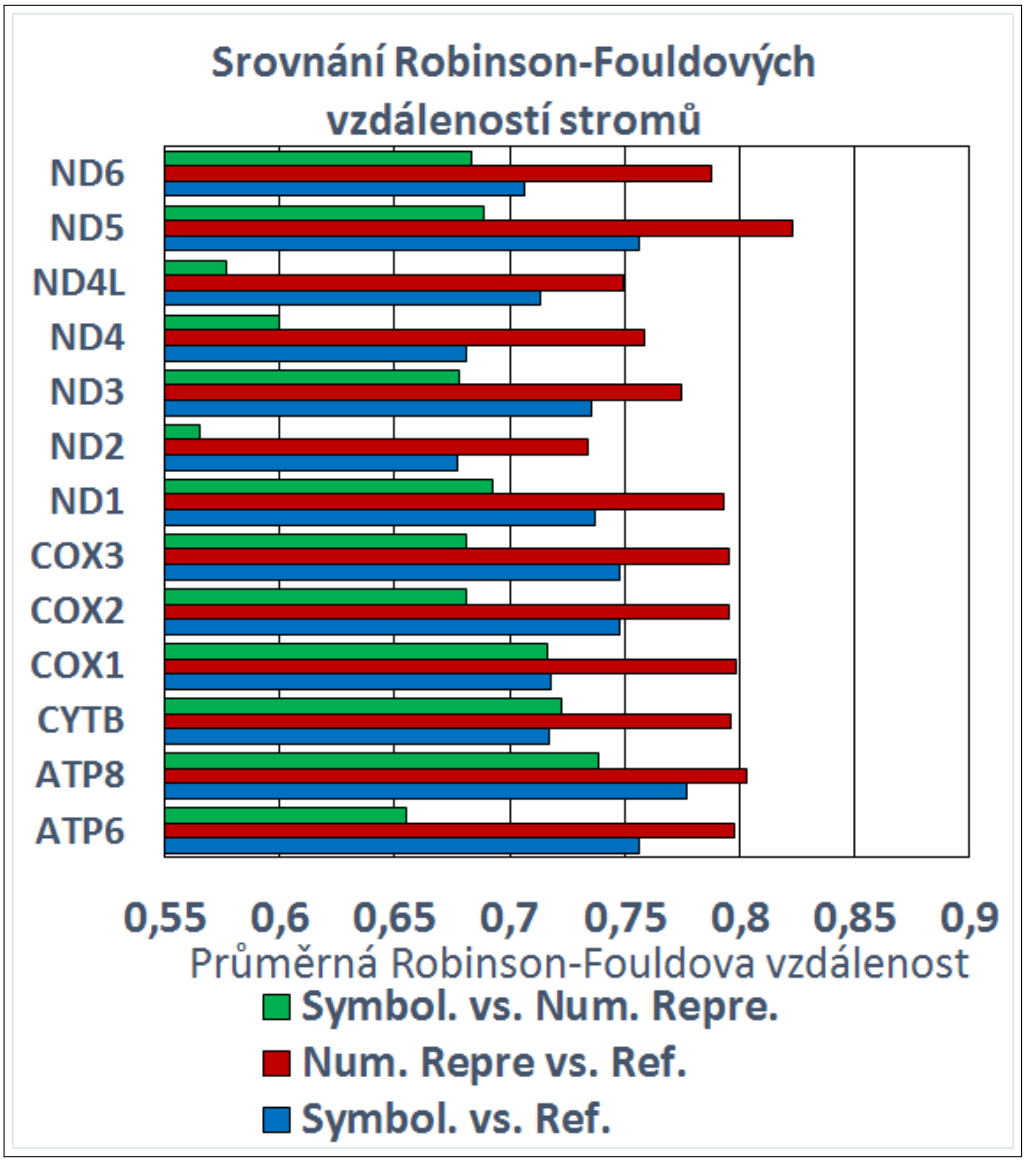
Obrázek 4.4: Srovnání RF vzdáleností reprezentací od standardní taxonomie při použití globální distanční matice.

4.6 Porovnání se symbolickou reprezentací

Nejlépe vyhodnocené reprezentace z předchozího testování jsem si vybral k porovnání s běžně používanými symbolickými reprezentacemi. Postup testování jsem nijak neměnil. Distanční matice pro zarovnané sekvence proteinů byla spočítána pomocí algoritmu Jukes-Cantor. Pro tvorbu stromů jsem zvolil metodu first-order, implementovanou v bioinformatickém toolboxu pro Matlab [28]. Výsledky testů, tedy průměrné hodnoty vzdáleností od standardního stromu, nalezneme v příloze této práce (viz. Tabulka A.2). Stejným způsobem byla změřena i vzdálenost stromů vytvořených nejvhodnější numerickou reprezentací a stromem vytvořeným ze symbolické reprezentace. Výsledné srovnání vzdáleností pro jednotlivé proteiny nalezneme na Obrázku 4.4. V grafu není zobrazena vzájemná vzdálenost stromů numerické a symbolické reprezentace, jelikož se její průměrné hodnoty pohybují pod hranicí 0,2 %. Touto metodikou vytvořené stromy jsou téměř identické pro numerickou i symbolickou reprezentaci.

Pro snížení výpočetních nároků na zpracování testů jsem převzorkován na stejnou vzdálenost (zarovnání v případě symbolické reprezentace) a výpočet distanční matice provedl vždy pro celý soubor všech testovaných organismů. Z této matice byly následně vybrány hodnoty vztahující se pouze k aktuálně zpracovávanému stromu. Vzdálenosti zpracovávané skupiny organismů tak mohly být zkreslené velikostí celého souboru.

Testování kvality jsem provedl pro nejvhodnější numerické reprezentace a symbolickou reprezentaci znovu. Algoritmus pro tento test upravím tak, aby se převzorkování a distanční matice počítala vždy pro každý konstruovaný strom zvlášť. V případě symbolické reprezentace provedeme pro každý strom zvlášť globální zarovnání a výpočet distanční matice. K zarovnání využijeme skórovací matici GONNET, která byla použita i při zarovnávání celého souboru. Metoda výpočtu vzdáleností a konstrukce stromu je stejná jako v předchozím testu. Výsledné průměrné vzdálenosti najdeme v příloze této práce (viz. Tabulka A.1). Z grafické reprezentace (viz. Obrázek 4.5) vidíme, že změna ve výpočtu neměla na numerické reprezentace téměř žádný vliv. U symbolické reprezentace pozorujeme výrazné zlepšení, a tato reprezentace se standardní taxonomii blíží více ve všech testovaných proteinech, než zvolené reprezentace numerické. Jako nejvhodnější protein se stále jeví protein ND2.



Obrázek 4.5: Srovnání RF vzdáleností reprezentací při použití výběrových distančních matic.

5. Závěr

Cílem práce bylo provést rešerši na téma numerické reprezentace proteinových sekvencí. Vybrané metody poté naprogramovat a navrhnout způsob testování jejich kvality. Testování provést na reálných datech a výsledky vyhodnotit a porovnat se standardní taxonomií.

V první části práce se věnuji různým způsobům numerických reprezentací proteinů. Reprezentace jsou rozděleny do podkapitol podle způsobu tvorby numerické reprezentace. Většina reprezentací je matematicky popsána a část i graficky prezentována. Z popsaných reprezentací jsem vybral pět, které jsem naprogramoval v prostředí Matlab.

V druhé části práce prezentuji naprogramované reprezentace a jejich použité modifikace. Popisuji zde také metodu testování kvality navržených reprezentací. Testování je založeno na porovnáváni Robinson-Fouldových vzdáleností fylogenetických stromů vytvořených z numerických reprezentací vůči referenčnímu stromu představujícího standardní taxonomii. Nejlépe vyhodnocené reprezentace porovnávám stejným způsobem také se symbolickou reprezentací stejného souboru proteinů.

Pro testování na reálných datech jsem zvolil soubor třinácti mitochondriálně kódovaných proteinů. Testování jsem provedl pro všechny naprogramované reprezentace a jejich modifikace, jakožto i pro všechny v práci uvedené metody výpočtu vzájemných vzdáleností sekvencí. Veškeré výsledky testů jsou uvedeny v Příloze A této práce. Vyhodnocení výsledků testů je graficky zpracováno a prezentováno na konci druhé části této práce.

Z výsledků testů je možno určit nejvhodnějším proteinem pro klasifikaci protein ND2, který vykazuje nejmenší odchylky vytvořených stromů od standardní taxonomie. Rovněž RF vzdálenost mezi symbolickou a nejvhodnější numerickou reprezentací je u tohoto proteinu nejmenší.

Pro každý gen se ukázala jiná vhodná kombinace numerické reprezentace a distanční metody. Mezi nejlépe vyhodnocenými reprezentacemi se nachází velká část reprezentací nekumulující numerické hodnoty.

Při porovnání stromů sestrojených numerickou a symbolickou reprezentací vůči standardní taxonomii se podařilo přiblížit se oběma metodami k podobným hodnotám RF vzdáleností od standardní taxonomie. Z celkového srovnání se stále jeví výhodnější použití symbolické reprezentace.

Seznam obrázků

Seznam tabulek

A. Výsledky testů

Tabulka A.1: Vyhodnocení průměrných RF vzdáleností při použití přímého výpočtu distančních matic.

Gen	RF vzdálenost	Symbol. vs. Ref.	Num. Repre vs. Ref.	Symbol. vs. Num. Repre.
ATP6	**mean**	0,7563	0,7978	0,6548
	std	0,0954	0,0901	0,1325
ATP8	**mean**	0,7772	0,8032	0,7387
	std	0,0896	0,0886	0,1095
CYTB	**mean**	0,7172	0,7960	0,7222
	std	0,0902	0,0880	0,0110
COX1	**mean**	0,7179	0,7988	0,7161
	std	0,0875	0,0887	0,1150
COX2	**mean**	0,7476	0,7952	0,6808
	std	0,0842	0,0769	0,1012
COX3	**mean**	0,7476	0,7952	0,6808
	std	0,0922	0,0764	0,0853
ND1	**mean**	0,7371	0,7928	0,6926
	std	0,0875	0,0756	0,1154
ND2	**mean**	0,6771	0,7337	0,5654
	std	0,0679	0,0742	0,0987
ND3	**mean**	0,7357	0,7750	0,6784
	std	0,0754	0,0803	0,0911
ND4	**mean**	0,6809	0,7588	0,6002
	std	0,0802	0,0746	0,0821
ND4L	**mean**	0,7136	0,7497	0,5771
	std	0,0722	0,0746	0,0954
ND5	**mean**	0,7566	0,8232	0,6889
	std	0,0748	0,0746	0,1357
ND6	**mean**	0,7064	0,7878	0,6831
	std	0,0766	0,0942	0,1277

Tabulka A.2: Vyhodnocení Robinson-Fouldových vzdáleností nejvhodnějších reprezentací vůči referenčnímu stromu.

Gen	1. num. repre.	2. num. repre.	3. num. repre.	Symbolická repre.	Symbolická vs. 1. num. repre.
ATP6 délka: 227 aa (207aa - 272aa)	0,7931 rep_hydrofil2 (okno=4) cityblock	0,7954 rep_eiip (nekumulovane) hamming	0,7981 rep_disoc1 (nekumulovane) jaccard	0,8827 seqneighjoin (firstorder) Jukes-Cantor	0,0016
ATP8 délka: 58 aa (50aa - 71aa)	0,7998 rep_disoc1 (nekumulovane) cosine	0,8102 rep_hydrofil2 (okno=4) cityblock	0,8108 rep_eiip (nekumulovane) cosine	0,8961 seqneighjoin (firstorder) Jukes-Cantor	0,0017
CYTB délka: 379 aa (284aa - 413aa)	0,7874 rep_eiip (nekumulovane) cityblock	0,7977 rep_hydrofil2 (okno=12) cityblock	0,7980 rep_hydrofil2 (okno=12)(detrend) cityblock	0,8829 seqneighjoin (firstorder) Jukes-Cantor	0,0017
COX1 délka: 517 aa (425aa - 543aa)	0,7919 rep_hydrofil2 (okno=30) cityblock	0,7983 rep_hydrofil2 (okno=30)(detrend) cityblock	0,7983 rep_eiip (nekumulovane) cityblock	0,8912 seqneighjoin (firstorder) Jukes-Cantor	0,0018
COX2 délka: 229 aa (166aa - 268aa)	0,7953 rep_hydrofil2 (okno=12) cityblock	0,7960 rep_hydrofil2 (okno=4) cityblock	0,7977 rep_eiip (nekumulovane) hamming	0,8696 seqneighjoin (firstorder) Jukes-Cantor	0,0018
COX3 délka: 261 aa (215aa - 279)	0,7830 rep_eiip (nekumulovane) jaccard	0,7859 rep_disoc1 (nekumulovane) hamming	0,7911 rep_hydrofil2 (okno=12) jaccard	0,8822 seqneighjoin (firstorder) Jukes-Cantor	0,0018
ND1 délka: 322 aa (311aa - 332aa)	0,7939 rep_eiip (nekumulovane) cityblock	0,7947 rep_disoc1 (nekumulovane) euclidean	0,8031 rep_hydrofil2 (okno=4)(detrend) cityblock	0,8757 seqneighjoin (firstorder) Jukes-Cantor	0,0018
ND2 délka: 347 aa (339aa - 369aa)	0,7246 rep_eiip (nekumulovane) cityblock	0,7247 rep_hydrofil2 (okno=4) cityblock	0,7317 rep_disoc1 (nekumulovane) cosine	0,8680 seqneighjoin (firstorder) Jukes-Cantor	0,0018
ND3 délka: 116 aa (103aa - 130aa)	0,7722 rep_disoc1 (nekumulovane) cityblock	0,7852 rep_eiip (nekumulovane) jaccard	0,7939 rep_hydrofil2 (okno=4) cityblock	0,8728 seqneighjoin (firstorder) Jukes-Cantor	0,0019
ND4 délka: 459 aa (445aa - 485aa)	0,7554 rep_disoc1 (nekumulovane) spearman	0,7613 rep_hydrofil2 (okno=4)(detrend) cityblock	0,7621 rep_eiip (nekumulovane) spearman	0,8573 seqneighjoin (firstorder) Jukes-Cantor	0,0019
ND4L délka: 98 aa (92aa - 101aa)	0,7546 rep_disoc1 (nekumulovane) jaccard	0,7646 rep_hydrofil2 (okno=4) cityblock	0,7648 rep_eiip (nekumulovane) jaccard	0,8711 seqneighjoin (firstorder) Jukes-Cantor	0,0017
ND5 délka: 608 aa (582aa - 635aa)	0,8041 rep_hydrofil2 (okno=12) euclidean	0,8067 rep_hydrofil2 (okno=12)(detrend) cosine	0,8164 rep_hydrofil2 (okno=4)(detrend) euclidean	0,8763 seqneighjoin (firstorder) Jukes-Cantor	0,0020
ND6 délka: 173 aa (156aa - 342aa)	0,7763 rep_eiip (nekumulovane) cosine	0,7764 rep_hydrofil2 (okno=12) cityblock	0,7769 rep_hydrofil2 (okno=4)(detrend) cityblock	0,8767 seqneighjoin (firstorder) Jukes-Cantor	0,0018

Tabulka A.3: Ohodnocení kvality reprezentací genu ATP6.

reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,9209	0,9052	0,8728	0,8987	0,8881	0,8704	0,8544	0,8691	0,9238
	std	0,0633	0,0664	0,0728	0,0699	0,0736	0,0807	0,0786	0,0767	0,0625
rep_eiip	**mean**	0,8709	0,8800	0,9948	0,8602	0,8597	0,8562	0,8702	0,9949	0,8773
(detrend)	**std**	0,0732	0,0754	0,0166	0,0761	0,0787	0,0793	0,0752	0,0165	0,0768
rep_eiip	**mean**	0,8014	0,8314	0,7954	0,8787	0,8024	0,8013	0,8043	0,8036	0,8049
(nekumulovane)	**std**	0,0861	0,0801	0,0848	0,0765	0,0847	0,0830	0,0881	0,0826	0,0868
rep_disoc1	**mean**	0,9082	0,8831	0,8843	0,8838	0,8808	0,8653	0,9617	0,8733	0,9104
	std	0,0643	0,0744	0,0708	0,0709	0,0748	0,0770	0,0434	0,0755	0,0638
rep_disoc1	**mean**	0,8711	0,8724	0,9954	0,8531	0,8570	0,8594	0,8658	0,9952	0,8766
(detrend)	**std**	0,0788	0,0742	0,0153	0,0788	0,0785	0,0791	0,0738	0,0167	0,0740
rep_disoc1	**mean**	0,8146	0,8406	0,7986	0,8756	0,8148	0,8107	0,8030	0,7981	0,7999
(nekumulovane)	**std**	0,0823	0,0804	0,0849	0,0770	0,0830	0,0863	0,0899	0,0804	0,0897
rep_hydrofil2	**mean**	0,8047	0,8143	0,8018	0,8603	0,8113	0,8092	0,8053	0,7943	0,7931
(okno=4)	**std**	0,0966	0,0828	0,0832	0,0790	0,0866	0,0868	0,0891	0,0913	0,0879
rep_hydrofil2	**mean**	0,8161	0,8124	0,9910	0,8491	0,8154	0,8140	0,8078	0,9893	0,8040
(okno=4)(detrend)	**std**	0,0855	0,0822	0,0225	0,0820	0,0841	0,0828	0,0891	0,0238	0,0844
rep_hydrofil2	**mean**	0,8197	0,8112	0,8146	0,8436	0,8223	0,8136	0,8216	0,8091	0,8074
(okno=12)	**std**	0,0891	0,0863	0,0838	0,0792	0,0860	0,0836	0,0821	0,0825	0,0849
rep_hydrofil2	**mean**	0,8134	0,8170	0,9938	0,8394	0,8159	0,8136	0,8218	0,9939	0,8043
(okno=12)(detrend)	**std**	0,0816	0,0834	0,0175	0,0809	0,0848	0,0851	0,0806	0,0181	0,0879
rep_hydrofil2	**mean**	0,8329	0,9974	0,8483	0,8462	0,8508	0,8453	0,8477	0,8540	0,8307
(okno=30)	**std**	0,0841	0,0116	0,0794	0,0826	0,0800	0,0804	0,0768	0,0807	0,0864
rep_hydrofil2	**mean**	0,8467	0,8438	0,9939	0,8407	0,8398	0,8383	0,8376	0,9946	0,8526
(okno=30)(detrend)	**std**	0,0815	0,0799	0,0178	0,0812	0,0797	0,0792	0,0839	0,0180	0,0773
rep_izoel	**mean**	0,9052	0,8917	0,8654	0,8754	0,8721	0,8611	0,9736	0,8699	0,9060
	std	0,0684	0,0694	0,0765	0,0739	0,0746	0,0767	0,0358	0,0741	0,0655
rep_izoel	**mean**	0,8671	0,8692	0,9950	0,8476	0,8556	0,8560	0,8530	0,9940	0,8700
(detrend)	**std**	0,0742	0,0759	0,0163	0,0799	0,0755	0,0793	0,0769	0,0183	0,0716
rep_izoel	**mean**	0,8974	0,8882	0,8716	0,8714	0,8701	0,8654	0,9676	0,8709	0,9119
(nekumulovane)	**std**	0,0676	0,0673	0,0763	0,0701	0,0798	0,0814	0,0380	0,0737	0,0616
rep_PHCurve	**mean**	0,9201	0,8968	0,9880	0,9060	0,8922	0,8932	0,9957	0,9956	0,9234
	std	0,0607	0,0690	0,0257	0,0674	0,0713	0,0681	0,0161	0,0155	0,0610
rep_PHCurve	**mean**	0,8794	0,8807	0,9951	0,8774	0,8824	0,8810	0,9123	0,9959	0,8799
(detrend)	**std**	0,0768	0,0760	0,0165	0,0732	0,0706	0,0735	0,0690	0,0150	0,0739

Tabulka A.4: Ohodnocení kvality reprezentací genu ATP8.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,8960	0,8863	0,9370	0,8844	0,8739	0,8648	0,8596	0,9370	0,8967
	std	0,0718	0,0708	0,0571	0,0727	0,0755	0,0777	0,0795	0,0566	0,0712
rep_eiip	**mean**	0,8768	0,8742	0,9933	0,8720	0,8691	0,8716	0,8828	0,9946	0,8752
(detrend)	**std**	0,0800	0,0722	0,0191	0,0722	0,0727	0,0766	0,0707	0,0169	0,0733
rep_eiip	**mean**	0,8121	0,8161	0,9154	0,8811	0,8108	0,8193	0,8244	0,9181	0,8221
(nekumulovane)	**std**	0,0849	0,0821	0,0645	0,0711	0,0867	0,0827	0,0796	0,0616	0,0838
rep_disoc1	**mean**	0,8648	0,8551	0,9400	0,8542	0,8439	0,8483	0,9311	0,9393	0,8708
	std	0,0716	0,0746	0,0558	0,0749	0,0779	0,0776	0,0570	0,0557	0,0744
rep_disoc1	**mean**	0,8494	0,8389	0,9944	0,8439	0,8398	0,8389	0,8387	0,9934	0,8503
(detrend)	**std**	0,0821	0,0833	0,0174	0,0840	0,0779	0,0798	0,0853	0,0193	0,0840
rep_disoc1	**mean**	0,8001	0,8086	0,9150	0,8602	0,7998	0,8072	0,8039	0,9133	0,8013
(nekumulovane)	**std**	0,0868	0,0842	0,0587	0,0809	0,0845	0,0872	0,0918	0,0614	0,0854
rep_hydrofil2	**mean**	0,8138	0,8252	0,9138	0,8494	0,8261	0,8271	0,8292	0,9162	0,8102
(okno=4)	**std**	0,0879	0,0865	0,0637	0,0817	0,0842	0,0850	0,0818	0,0680	0,0834
rep_hydrofil2	**mean**	0,8248	0,8261	0,9953	0,8613	0,8247	0,8266	0,8228	0,9904	0,8222
(okno=4)(detrend)	**std**	0,0832	0,0835	0,0170	0,0839	0,0848	0,0861	0,0832	0,0219	0,0823
rep_hydrofil2	**mean**	0,8404	0,8546	0,9320	0,8526	0,8699	0,8476	0,8758	0,9350	0,8510
(okno=12)	**std**	0,0843	0,0819	0,0597	0,0790	0,0787	0,0772	0,0754	0,0588	0,0821
rep_hydrofil2	**mean**	0,8622	0,8579	0,9941	0,8616	0,8639	0,8540	0,8660	0,9928	0,8652
(okno=12)(detrend)	**std**	0,0765	0,0772	0,0175	0,0745	0,0760	0,0777	0,0831	0,0197	0,0811
rep_hydrofil2	**mean**	0,8949	0,9969	0,9507	0,8773	0,9029	0,8947	0,9071	0,9582	0,8961
(okno=30)	**std**	0,0694	0,0137	0,0495	0,0749	0,0660	0,0698	0,0656	0,0473	0,0690
rep_hydrofil2	**mean**	0,8763	0,8807	0,9929	0,8853	0,8754	0,8750	0,8991	0,9928	0,8802
(okno=30)(detrend)	**std**	0,0712	0,0739	0,0186	0,0728	0,0739	0,0735	0,0699	0,0197	0,0771
rep_izoel	**mean**	0,8707	0,8639	0,9389	0,8677	0,8529	0,8472	0,9342	0,9409	0,8680
	std	0,0757	0,0760	0,0559	0,0749	0,0771	0,0813	0,0550	0,0532	0,0748
rep_izoel	**mean**	0,8546	0,8572	0,9930	0,8569	0,8551	0,8571	0,8631	0,9936	0,8638
(detrend)	**std**	0,0779	0,0772	0,0201	0,0820	0,0828	0,0794	0,0804	0,0182	0,0742
rep_izoel	**mean**	0,8677	0,8653	0,9426	0,8574	0,8508	0,8489	0,9384	0,9417	0,8680
(nekumulovane)	**std**	0,0765	0,0758	0,0536	0,0773	0,0799	0,0827	0,0539	0,0536	0,0738
rep_PHCurve	**mean**	0,8847	0,8989	0,9944	0,8771	0,8703	0,8727	0,9447	0,9912	0,8966
	std	0,0719	0,0670	0,0167	0,0732	0,0741	0,0738	0,0506	0,0218	0,0630
rep_PHCurve	**mean**	0,8712	0,8926	0,9932	0,8642	0,8713	0,8687	0,9031	0,9949	0,8744
(detrend)	**std**	0,0759	0,0678	0,0185	0,0733	0,0733	0,0745	0,0656	0,0165	0,0734

Tabulka A.5: Ohodnocení kvality reprezentací genu CYTB.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,8984	0,8890	0,9013	0,8898	0,8588	0,8479	0,8404	0,9072	0,9018
	std	0,0691	0,0732	0,0713	0,0687	0,0839	0,0875	0,0793	0,0716	0,0688
rep_eiip	**mean**	0,8508	0,8484	0,9948	0,8474	0,8493	0,8498	0,8541	0,9959	0,8583
(detrend)	**std**	0,0786	0,0834	0,0166	0,0836	0,0824	0,0795	0,0824	0,0146	0,0818
rep_eiip	**mean**	0,7942	0,8169	0,8018	0,8652	0,8039	0,8042	0,8011	0,8034	0,7874
(nekumulovane)	**std**	0,0903	0,0841	0,0902	0,0796	0,0873	0,0805	0,0898	0,0901	0,0913
rep_disoc1	**mean**	0,9161	0,8757	0,9017	0,9027	0,8782	0,8613	0,9874	0,9094	0,9139
	std	0,0593	0,0736	0,0713	0,0670	0,0799	0,0792	0,0260	0,0695	0,0651
rep_disoc1	**mean**	0,8626	0,8590	0,9948	0,8591	0,8623	0,8563	0,8594	0,9958	0,8714
(detrend)	**std**	0,0808	0,0802	0,0166	0,0807	0,0780	0,0776	0,0770	0,0152	0,0778
rep_disoc1	**mean**	0,8086	0,8296	0,8038	0,8691	0,8169	0,8182	0,8193	0,8008	0,8121
(nekumulovane)	**std**	0,0894	0,0782	0,0900	0,0791	0,0881	0,0757	0,0854	0,0933	0,0879
rep_hydrofil2	**mean**	0,8086	0,8098	0,8081	0,8626	0,8096	0,8026	0,8066	0,8011	0,8010
(okno=4)	**std**	0,0868	0,0854	0,0871	0,0772	0,0857	0,0946	0,0860	0,0934	0,0895
rep_hydrofil2	**mean**	0,8016	0,8053	0,9943	0,8649	0,8094	0,8142	0,8104	0,9944	0,8027
(okno=4)(detrend)	**std**	0,0913	0,0876	0,0172	0,0755	0,0911	0,0886	0,0877	0,0167	0,0883
rep_hydrofil2	**mean**	0,8133	0,9973	0,8052	0,8629	0,8108	0,8060	0,8108	0,8194	0,7977
(okno=12)	**std**	0,0874	0,0119	0,0868	0,0777	0,0870	0,0890	0,0819	0,0840	0,0870
rep_hydrofil2	**mean**	0,8187	0,8057	0,9947	0,8581	0,8042	0,8113	0,8064	0,9943	0,7980
(okno=12)(detrend)	**std**	0,0816	0,0884	0,0171	0,0819	0,0916	0,0874	0,0867	0,0176	0,0810
rep_hydrofil2	**mean**	0,8323	0,9966	0,8372	0,8569	0,8320	0,8373	0,8432	0,8411	0,8216
(okno=30)	**std**	0,0802	0,0139	0,0817	0,0786	0,0864	0,0797	0,0816	0,0862	0,0844
rep_hydrofil2	**mean**	0,8324	0,8312	0,9950	0,8556	0,8404	0,8424	0,8366	0,9951	0,8259
(okno=30)(detrend)	**std**	0,0820	0,0856	0,0163	0,0809	0,0874	0,0844	0,0822	0,0161	0,0872
rep_izoel	**mean**	0,8982	0,8747	0,9004	0,8792	0,8701	0,8506	0,9908	0,8988	0,8877
	std	0,0695	0,0730	0,0688	0,0737	0,0711	0,0768	0,0230	0,0670	0,0724
rep_izoel	**mean**	0,8541	0,8584	0,9942	0,8554	0,8537	0,8456	0,8573	0,9952	0,8528
(detrend)	**std**	0,0798	0,0805	0,0173	0,0798	0,0761	0,0807	0,0770	0,0156	0,0846
rep_izoel	**mean**	0,8922	0,8811	0,8933	0,8779	0,8613	0,8441	0,9910	0,9030	0,8976
(nekumulovane)	**std**	0,0733	0,0719	0,0660	0,0730	0,0748	0,0822	0,0222	0,0642	0,0672
rep_PHCurve	**mean**	0,9334	0,9017	0,9809	0,9286	0,9170	0,9146	0,9903	0,9946	0,9387
	std	0,0582	0,0671	0,0315	0,0595	0,0679	0,0665	0,0219	0,0176	0,0543
rep_PHCurve	**mean**	0,9080	0,9040	0,9954	0,9069	0,8868	0,8889	0,9407	0,9953	0,9104
(detrend)	**std**	0,0688	0,0687	0,0153	0,0657	0,0739	0,0739	0,0572	0,0158	0,0660

Tabulka A.6: Ohodnocení kvality reprezentací genu COX1.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,8752	0,8644	0,8761	0,8510	0,8377	0,8240	0,8163	0,8723	0,8737
	std	0,0731	0,0792	0,0704	0,0773	0,0864	0,0837	0,0867	0,0686	0,0774
rep_eiip	**mean**	0,8252	0,8197	0,9938	0,8274	0,8161	0,8162	0,8127	0,9919	0,8189
(detrend)	**std**	0,0851	0,0866	0,0186	0,0842	0,0866	0,0844	0,0855	0,0206	0,0860
rep_eiip	**mean**	0,8046	0,8079	0,8090	0,8551	0,8086	0,7997	0,8017	0,8068	0,7983
(nekumulovane)	**std**	0,0913	0,0864	0,0885	0,0773	0,0872	0,0909	0,0889	0,0873	0,0842
rep_disoc1	**mean**	0,8804	0,8614	0,8778	0,8679	0,8580	0,8496	0,9472	0,8800	0,8776
	std	0,0739	0,0763	0,0739	0,0786	0,0799	0,0776	0,0476	0,0731	0,0712
rep_disoc1	**mean**	0,8470	0,8457	0,9933	0,8481	0,8463	0,8446	0,8441	0,9949	0,8496
(detrend)	**std**	0,0823	0,0783	0,0197	0,0739	0,0805	0,0765	0,0803	0,0165	0,0780
rep_disoc1	**mean**	0,8394	0,8244	0,8009	0,8516	0,8337	0,8380	0,8321	0,8060	0,8169
(nekumulovane)	**std**	0,0827	0,0842	0,0869	0,0809	0,0811	0,0805	0,0808	0,0875	0,0882
rep_hydrofil2	**mean**	0,8142	0,8237	0,8060	0,8506	0,8094	0,8037	0,8146	0,8089	0,8024
(okno=4)	**std**	0,0873	0,0826	0,0835	0,0764	0,0855	0,0878	0,0846	0,0837	0,0892
rep_hydrofil2	**mean**	0,8074	0,8152	0,9940	0,8487	0,8058	0,8146	0,8119	0,9914	0,8019
(okno=4)(detrend)	**std**	0,0873	0,0846	0,0180	0,0824	0,0868	0,0866	0,0868	0,0216	0,0888
rep_hydrofil2	**mean**	0,8120	0,9969	0,8119	0,8433	0,8124	0,8116	0,8084	0,8100	0,8017
(okno=12)	**std**	0,0852	0,0137	0,0816	0,0803	0,0830	0,0857	0,0899	0,0837	0,0834
rep_hydrofil2	**mean**	0,8076	0,8106	0,9929	0,8466	0,8074	0,8182	0,8082	0,9940	0,8081
(okno=12)(detrend)	**std**	0,0879	0,0811	0,0196	0,0809	0,0851	0,0873	0,0869	0,0176	0,0869
rep_hydrofil2	**mean**	0,8054	0,9953	0,8171	0,8281	0,8016	0,8116	0,8048	0,8130	0,7919
(okno=30)	**std**	0,0925	0,0166	0,0866	0,0855	0,0912	0,0903	0,0903	0,0883	0,0915
rep_hydrofil2	**mean**	0,8113	0,8007	0,9921	0,8357	0,8059	0,8141	0,8052	0,9942	0,7983
(okno=30)(detrend)	**std**	0,0869	0,0873	0,0209	0,0884	0,0922	0,0861	0,0874	0,0177	0,0923
rep_izoel	**mean**	0,8694	0,8652	0,8758	0,8580	0,8568	0,8434	0,9586	0,8773	0,8718
	std	0,0777	0,0754	0,0749	0,0800	0,0794	0,0800	0,0492	0,0763	0,0730
rep_izoel	**mean**	0,8532	0,8569	0,9922	0,8423	0,8519	0,8517	0,8482	0,9943	0,8509
(detrend)	**std**	0,0746	0,0745	0,0211	0,0820	0,0789	0,0769	0,0780	0,0176	0,0756
rep_izoel	**mean**	0,8628	0,8680	0,8724	0,8557	0,8498	0,8440	0,9568	0,8716	0,8668
(nekumulovane)	**std**	0,0739	0,0761	0,0682	0,0811	0,0782	0,0783	0,0460	0,0740	0,0794
rep_PHCurve	**mean**	0,9078	0,8901	0,9899	0,8872	0,8827	0,8722	0,9546	0,9940	0,9110
	std	0,0651	0,0705	0,0242	0,0730	0,0737	0,0738	0,0458	0,0183	0,0656
rep_PHCurve	**mean**	0,8826	0,8837	0,9939	0,8749	0,8666	0,8644	0,9079	0,9953	0,8887
(detrend)	**std**	0,0743	0,0731	0,0188	0,0725	0,0752	0,0737	0,0655	0,0162	0,0742

Tabulka A.7: Ohodnocení kvality reprezentací genu COX2.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,8751	0,8640	0,8856	0,8597	0,8368	0,8276	0,8171	0,8772	0,8757
	std	0,0769	0,0770	0,0726	0,0801	0,0818	0,0860	0,0901	0,0740	0,0753
rep_eiip (detrend)	**mean**	0,8190	0,8189	0,9924	0,8264	0,8197	0,8200	0,8244	0,9936	0,8296
	std	0,0829	0,0849	0,0200	0,0835	0,0902	0,0836	0,0830	0,0185	0,0863
rep_eiip (nekumulovane)	**mean**	0,8070	0,8201	0,7977	0,8403	0,8072	0,8169	0,8136	0,8089	0,8028
	std	0,0891	0,0798	0,0931	0,0851	0,0890	0,0869	0,0855	0,0825	0,0934
rep_disoc1	**mean**	0,8784	0,8482	0,8821	0,8539	0,8508	0,8489	0,9666	0,8737	0,8748
	std	0,0735	0,0849	0,0708	0,0760	0,0826	0,0798	0,0433	0,0708	0,0734
rep_disoc1 (detrend)	**mean**	0,8509	0,8558	0,9923	0,8429	0,8418	0,8481	0,8336	0,9907	0,8583
	std	0,0808	0,0788	0,0198	0,0806	0,0821	0,0786	0,0829	0,0228	0,0726
rep_disoc1 (nekumulovane)	**mean**	0,8143	0,8228	0,8037	0,8469	0,8107	0,8112	0,8114	0,7994	0,8086
	std	0,0851	0,0824	0,0907	0,0833	0,0904	0,0843	0,0896	0,0855	0,0913
rep_hydrofil2 (okno=4)	**mean**	0,8019	0,8137	0,7988	0,8381	0,8113	0,8107	0,8192	0,8009	0,7960
	std	0,0877	0,0893	0,0859	0,0853	0,0851	0,0921	0,0887	0,0897	0,0866
rep_hydrofil2 (okno=4)(detrend)	**mean**	0,8039	0,8143	0,9924	0,8440	0,8090	0,8123	0,8149	0,9919	0,8017
	std	0,0902	0,0841	0,0197	0,0807	0,0860	0,0845	0,0843	0,0209	0,0922
rep_hydrofil2 (okno=12)	**mean**	0,8110	0,8128	0,8051	0,8366	0,8080	0,8101	0,8109	0,8072	0,7953
	std	0,0844	0,0884	0,0839	0,0815	0,0882	0,0863	0,0886	0,0903	0,0873
rep_hydrofil2 (okno=12)(detrend)	**mean**	0,8116	0,8160	0,9907	0,8292	0,8063	0,8066	0,8171	0,9937	0,8093
	std	0,0897	0,0848	0,0228	0,0800	0,0870	0,0848	0,0848	0,0197	0,0875
rep_hydrofil2 (okno=30)	**mean**	0,8208	0,9959	0,8383	0,8136	0,8187	0,8237	0,8242	0,8364	0,8104
	std	0,0912	0,0146	0,0839	0,0847	0,0908	0,0856	0,0870	0,0791	0,0882
rep_hydrofil2 (okno=30)(detrend)	**mean**	0,8212	0,8263	0,9911	0,8350	0,8071	0,8180	0,8303	0,9910	0,8223
	std	0,0838	0,0850	0,0219	0,0828	0,0887	0,0854	0,0841	0,0225	0,0831
rep_izoel	**mean**	0,8727	0,8669	0,8774	0,8562	0,8494	0,8496	0,9679	0,8732	0,8678
	std	0,0770	0,0781	0,0740	0,0795	0,0809	0,0792	0,0409	0,0765	0,0782
rep_izoel (detrend)	**mean**	0,8598	0,8458	0,9921	0,8471	0,8498	0,8442	0,8527	0,9922	0,8510
	std	0,0813	0,0818	0,0206	0,0816	0,0781	0,0808	0,0774	0,0202	0,0789
rep_izoel (nekumulovane)	**mean**	0,8662	0,8601	0,8773	0,8612	0,8517	0,8459	0,9663	0,8789	0,8689
	std	0,0783	0,0779	0,0744	0,0790	0,0822	0,0776	0,0419	0,0744	0,0769
rep_PHCurve	**mean**	0,8917	0,8751	0,9830	0,8864	0,8672	0,8613	0,9754	0,9908	0,8916
	std	0,0709	0,0698	0,0313	0,0735	0,0771	0,0783	0,0355	0,0230	0,0768
rep_PHCurve (detrend)	**mean**	0,8741	0,8681	0,9912	0,8616	0,8571	0,8641	0,8887	0,9907	0,8739
	std	0,0782	0,0763	0,0226	0,0769	0,0783	0,0783	0,0736	0,0219	0,0735

Tabulka A.8: Ohodnocení kvality reprezentací genu COX3.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,9109	0,8843	0,8748	0,8823	0,8774	0,8673	0,8591	0,8717	0,9054
	std	0,0653	0,0776	0,0745	0,0706	0,0758	0,0798	0,0782	0,0753	0,0748
rep_eiip	**mean**	0,8658	0,8688	0,9941	0,8451	0,8630	0,8669	0,8702	0,9956	0,8702
(detrend)	**std**	0,0778	0,0800	0,0175	0,0810	0,0775	0,0751	0,0789	0,0151	0,0752
rep_eiip	**mean**	0,7968	0,8447	0,7840	0,8857	0,7994	0,7944	0,7917	0,7830	0,7997
(nekumulovane)	**std**	0,0886	0,0848	0,0902	0,0749	0,0887	0,0856	0,0926	0,0884	0,0916
rep_disoc1	**mean**	0,9153	0,8983	0,8814	0,8530	0,8776	0,8722	0,9960	0,8848	0,9161
	std	0,0612	0,6840	0,0744	0,0779	0,0779	0,0777	0,0148	0,0752	0,0636
rep_disoc1	**mean**	0,8722	0,8712	0,9938	0,8499	0,8689	0,8719	0,8659	0,9949	0,8714
(detrend)	**std**	0,0773	0,0771	0,0192	0,0836	0,0756	0,0761	0,0753	0,0165	0,0789
rep_disoc1	**mean**	0,8041	0,8458	0,7859	0,8733	0,8038	0,8037	0,7967	0,7916	0,7950
(nekumulovane)	**std**	0,0885	0,0793	0,0921	0,0738	0,0870	0,0869	0,0905	0,0922	0,0956
rep_hydrofil2	**mean**	0,8286	0,8344	0,7957	0,8836	0,8279	0,8239	0,8229	0,7921	0,8110
(okno=4)	**std**	0,0835	0,0854	0,0878	0,0743	0,0825	0,0828	0,0860	0,0888	0,0850
rep_hydrofil2	**mean**	0,8344	0,8328	0,9850	0,8758	0,8253	0,8136	0,8293	0,9862	0,8127
(okno=4)(detrend)	**std**	0,0851	0,0852	0,0273	0,0772	0,0831	0,0840	0,0815	0,0274	0,0855
rep_hydrofil2	**mean**	0,8358	0,9960	0,7947	0,8744	0,8441	0,8404	0,8458	0,7911	0,8177
(okno=12)	**std**	0,0830	0,0148	0,0938	0,0801	0,0798	0,0792	0,0796	0,0911	0,0880
rep_hydrofil2	**mean**	0,8359	0,8369	0,9882	0,8732	0,8396	0,8364	0,8444	0,9896	0,8236
(okno=12)(detrend)	**std**	0,0830	0,0821	0,0243	0,0753	0,0848	0,0852	0,0823	0,0249	0,0829
rep_hydrofil2	**mean**	0,8656	0,9970	0,8452	0,8651	0,8696	0,8781	0,8693	0,8443	0,8513
(okno=30)	**std**	0,0774	0,0126	0,0812	0,0823	0,0757	0,0745	0,0783	0,0825	0,0811
rep_hydrofil2	**mean**	0,8699	0,8696	0,9930	0,8620	0,8690	0,8657	0,8763	0,9920	0,8703
(okno=30)(detrend)	**std**	0,0744	0,0748	0,0194	0,0765	0,0774	0,0810	0,0761	0,0208	0,0764
rep_izoel	**mean**	0,8940	0,8720	0,8651	0,8828	0,8729	0,8737	0,9957	0,8703	0,8944
	std	0,0687	0,0758	0,0760	0,0735	0,0775	0,0756	0,0149	0,0800	0,0699
rep_izoel	**mean**	0,8761	0,8747	0,9940	0,8614	0,8749	0,8783	0,8823	0,9942	0,8788
(detrend)	**std**	0,0787	0,0742	0,0173	0,0825	0,0713	0,0773	0,0755	0,0177	0,0761
rep_izoel	**mean**	0,8989	0,8698	0,8754	0,8871	0,8769	0,8743	0,9946	0,8756	0,8897
(nekumulovane)	**std**	0,0700	0,0754	0,0786	0,0748	0,0770	0,0772	0,0165	0,0766	0,0720
rep_PHCurve	**mean**	0,9257	0,8902	0,9906	0,9268	0,8997	0,9074	0,9947	0,9943	0,9268
	std	0,0624	0,0725	0,0231	0,0617	0,0711	0,0712	0,0171	0,0176	0,0650
rep_PHCurve	**mean**	0,9091	0,9010	0,9939	0,9038	0,8969	0,9011	0,9114	0,9943	0,9093
(detrend)	**std**	0,0674	0,0669	0,0184	0,0706	0,0733	0,0688	0,0681	0,0168	0,0691

Tabulka A.9: Ohodnocení kvality reprezentací genu ND1.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,9058	0,8872	0,9472	0,8827	0,8784	0,8506	0,8092	0,9476	0,9026
	std	0,0686	0,0762	0,0498	0,0742	0,0725	0,0794	0,0865	0,0521	0,0671
rep_eiip (detrend)	**mean**	0,8584	0,8516	0,9952	0,8433	0,8400	0,8404	0,8518	0,9949	0,8661
	std	0,0773	0,0837	0,0160	0,0777	0,0772	0,0842	0,0775	0,0161	0,0740
rep_eiip (nekumulovane)	**mean**	0,8009	0,8081	0,8036	0,8564	0,7971	0,7970	0,8009	0,8078	0,7939
	std	0,0836	0,0851	0,0606	0,0795	0,0922	0,0924	0,0856	0,0851	0,0889
rep_disoc1	**mean**	0,8916	0,8866	0,9470	0,8602	0,8588	0,8449	0,9548	0,9466	0,8896
	std	0,0710	0,0703	0,0513	0,0783	0,0791	0,0799	0,0468	0,0523	0,0724
rep_disoc1 (detrend)	**mean**	0,8603	0,8641	0,9946	0,8482	0,8489	0,8520	0,8594	0,9941	0,8646
	std	0,0814	0,0816	0,0173	0,0806	0,0785	0,0787	0,0774	0,0188	0,0795
rep_disoc1 (nekumulovane)	**mean**	0,7947	0,8358	0,8148	0,8570	0,7994	0,8002	0,8062	0,8110	0,7996
	std	0,0892	0,0799	0,0876	0,0819	0,0873	0,0872	0,0884	0,0869	0,0904
rep_hydrofil2 (okno=4)	**mean**	0,8173	0,8202	0,8212	0,8426	0,8133	0,8100	0,8142	0,8218	0,8044
	std	0,0895	0,0869	0,0806	0,0796	0,0898	0,0852	0,0887	0,0820	0,0881
rep_hydrofil2 (okno=4)(detrend)	**mean**	0,8059	0,8221	0,9943	0,8471	0,8089	0,8072	0,8173	0,9950	0,8031
	std	0,0846	0,0857	0,0168	0,0849	0,0907	0,0902	0,0876	0,0163	0,0934
rep_hydrofil2 (okno=12)	**mean**	0,8053	0,9960	0,8191	0,8389	0,8230	0,8120	0,8282	0,8243	0,8091
	std	0,0874	0,0148	0,0844	0,0829	0,0773	0,0901	0,0862	0,0829	0,0805
rep_hydrofil2 (okno=12)(detrend)	**mean**	0,8146	0,8201	0,9938	0,8508	0,8213	0,8188	0,8212	0,9938	0,8081
	std	0,0864	0,0833	0,0189	0,0844	0,0871	0,0875	0,0875	0,0189	0,0851
rep_hydrofil2 (okno=30)	**mean**	0,8328	0,9957	0,8429	0,8386	0,8413	0,8408	0,8502	0,8429	0,8217
	std	0,0827	0,0153	0,0822	0,0828	0,0813	0,0859	0,0802	0,0847	0,0846
rep_hydrofil2 (okno=30)(detrend)	**mean**	0,8377	0,8392	0,9944	0,8432	0,8356	0,8392	0,8390	0,9954	0,8304
	std	0,0807	0,0822	0,0171	0,0811	0,0863	0,0841	0,0764	0,0153	0,0822
rep_izoel	**mean**	0,8904	0,8778	0,9514	0,8751	0,8746	0,8542	0,9552	0,9537	0,8917
	std	0,0701	0,0754	0,0480	0,0747	0,0761	0,0788	0,0439	0,0493	0,0707
rep_izoel (detrend)	**mean**	0,8670	0,8724	0,9950	0,8510	0,8682	0,8586	0,8517	0,9944	0,8707
	std	0,0772	0,0719	0,0167	0,0791	0,0764	0,0790	0,0780	0,0171	0,0745
rep_izoel (nekumulovane)	**mean**	0,8913	0,8798	0,9488	0,8679	0,8734	0,8584	0,9564	0,9529	0,8896
	std	0,0665	0,0746	0,0512	0,0715	0,0724	0,0820	0,0443	0,0481	0,0719
rep_PHCurve	**mean**	0,9028	0,8973	0,9928	0,8933	0,8902	0,8873	0,9630	0,9952	0,8970
	std	0,0694	0,0696	0,0194	0,0740	0,0700	0,0706	0,0385	0,0167	0,0667
rep_PHCurve (detrend)	**mean**	0,8841	0,8851	0,9943	0,8782	0,8841	0,8880	0,9243	0,9943	0,8871
	std	0,0738	0,0748	0,0172	0,0779	0,0722	0,0720	0,0625	0,0172	0,0761

Tabulka A.10: Ohodnocení kvality reprezentací genu ND2.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,9040	0,8957	0,9120	0,8880	0,8786	0,8600	0,7708	0,9091	0,9088
	std	0,0731	0,0675	0,0601	0,0751	0,0738	0,0788	0,0927	0,0622	0,0633
rep_eiip (detrend)	**mean**	0,8571	0,8548	0,9961	0,8441	0,8534	0,8573	0,8600	0,9956	0,8570
	std	0,0782	0,0800	0,0150	0,0817	0,0816	0,0775	0,0733	0,0151	0,0770
rep_eiip (nekumulovane)	**mean**	0,7377	0,7428	0,7814	0,8264	0,7340	0,7274	0,7326	0,7898	0,7246
	std	0,0977	0,0965	0,0923	0,0856	0,0959	0,1043	0,0983	0,0929	0,0964
rep_disoc1	**mean**	0,9014	0,8930	0,9171	0,8789	0,8679	0,8586	0,9590	0,9120	0,9033
	std	0,0702	0,0718	0,0624	0,0747	0,0728	0,0800	0,0436	0,0643	0,0703
rep_disoc1 (detrend)	**mean**	0,8617	0,8627	0,9957	0,8574	0,8644	0,8577	0,8559	0,9952	0,8684
	std	0,0778	0,0765	0,0157	0,0747	0,0747	0,0755	0,0796	0,0156	0,0788
rep_disoc1 (nekumulovane)	**mean**	0,7444	0,7552	0,7836	0,8254	0,7317	0,7409	0,7379	0,7879	0,7389
	std	0,0928	0,0946	0,0905	0,0880	0,0954	0,0908	0,0911	0,0946	0,0992
rep_hydrofil2 (okno=4)	**mean**	0,7408	0,7409	0,7897	0,8093	0,7352	0,7338	0,7327	0,7952	0,7247
	std	0,0948	0,0937	0,0906	0,0871	0,0962	0,0919	0,0983	0,0875	0,0984
rep_hydrofil2 (okno=4)(detrend)	**mean**	0,7427	0,7361	0,9952	0,8154	0,7392	0,7394	0,7386	0,9959	0,7346
	std	0,1037	0,0964	0,0160	0,0856	0,0971	0,0917	0,0966	0,0146	0,1023
rep_hydrofil2 (okno=12)	**mean**	0,7688	0,9971	0,8317	0,8263	0,7700	0,7672	0,7704	0,8404	0,7587
	std	0,0916	0,0128	0,0810	0,0852	0,0924	0,0958	0,0875	0,0826	0,0917
rep_hydrofil2 (okno=12)(detrend)	**mean**	0,7566	0,7624	0,9953	0,8121	0,7649	0,7682	0,7661	0,9959	0,7692
	std	0,0982	0,0955	0,0154	0,0874	0,0940	0,0926	0,0945	0,0146	0,0903
rep_hydrofil2 (okno=30)	**mean**	0,8150	0,9961	0,8836	0,8324	0,8211	0,8110	0,8131	0,8824	0,8123
	std	0,0885	0,0146	0,0723	0,0850	0,0871	0,0858	0,0851	0,0694	0,0892
rep_hydrofil2 (okno=30)(detrend)	**mean**	0,8166	0,8136	0,9964	0,8259	0,8173	0,8140	0,8191	0,9958	0,8070
	std	0,0870	0,0913	0,0141	0,0868	0,0880	0,0850	0,0873	0,0147	0,0887
rep_izoel	**mean**	0,8891	0,8791	0,9131	0,8509	0,8513	0,8532	0,9561	0,9132	0,8919
	std	0,0720	0,0726	0,0665	0,0823	0,0807	0,0824	0,0439	0,0636	0,0721
rep_izoel (detrend)	**mean**	0,8462	0,8468	0,9953	0,8137	0,8541	0,8511	0,8404	0,9944	0,8656
	std	0,0802	0,0807	0,0162	0,0947	0,0783	0,0846	0,0786	0,0174	0,0780
rep_izoel (nekumulovane)	**mean**	0,8913	0,8699	0,9103	0,8454	0,8574	0,8509	0,9579	0,9158	0,8929
	std	0,0722	0,0778	0,0633	0,0794	0,0825	0,0801	0,0414	0,0621	0,0727
rep_PHCurve	**mean**	0,9061	0,8977	0,9786	0,8890	0,8843	0,8722	0,9853	0,9958	0,9072
	std	0,0679	0,0708	0,0332	0,0676	0,0745	0,0772	0,0285	0,0152	0,0676
rep_PHCurve (detrend)	**mean**	0,8730	0,8686	0,9946	0,8653	0,8942	0,8989	0,9099	0,9952	0,8761
	std	0,0813	0,0767	0,0165	0,0777	0,0693	0,0693	0,0712	0,0171	0,0777

Tabulka A.11: Ohodnocení kvality reprezentací genu ND3.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,9056	0,9013	0,9178	0,8860	0,8869	0,8726	0,8868	0,9183	0,9113
	std	0,0686	0,0693	0,0592	0,0763	0,0688	0,0772	0,0750	0,0645	0,0658
rep_eiip	**mean**	0,8737	0,8726	0,9924	0,8652	0,8580	0,8704	0,8681	0,9943	0,8748
(detrend)	**std**	0,0732	0,0780	0,0200	0,0770	0,0798	0,0774	0,0799	0,0176	0,0738
rep_eiip	**mean**	0,8002	0,8096	0,7896	0,8697	0,8011	0,7928	0,7983	0,7852	0,7853
(nekumulovane)	**std**	0,0870	0,0849	0,0908	0,0777	0,0899	0,0860	0,0858	0,0823	0,0943
rep_disoc1	**mean**	0,9120	0,9083	0,9284	0,8771	0,8671	0,8508	0,9656	0,9260	0,9184
	std	0,0668	0,0642	0,0587	0,0760	0,0777	0,0771	0,0404	0,0587	0,0655
rep_disoc1	**mean**	0,8652	0,8662	0,9923	0,8629	0,8607	0,8618	0,8636	0,9934	0,8768
(detrend)	**std**	0,0810	0,0767	0,0204	0,0722	0,0774	0,0797	0,0758	0,0199	0,0759
rep_disoc1	**mean**	0,7906	0,8031	0,7871	0,8748	0,7920	0,7870	0,7934	0,7908	0,7722
(nekumulovane)	**std**	0,0939	0,0919	0,0866	0,0736	0,0919	0,0903	0,0872	0,0871	0,0910
rep_hydrofil2	**mean**	0,8108	0,8133	0,8093	0,8507	0,8050	0,8026	0,8098	0,8056	0,7939
(okno=4)	**std**	0,0854	0,0845	0,0853	0,0806	0,0917	0,0932	0,0888	0,0831	0,0889
rep_hydrofil2	**mean**	0,7969	0,8122	0,9916	0,8424	0,7956	0,8042	0,8139	0,9932	0,8074
(okno=4)(detrend)	**std**	0,0895	0,0816	0,0209	0,0791	0,0926	0,0846	0,0895	0,0189	0,0870
rep_hydrofil2	**mean**	0,8377	0,8522	0,8331	0,8641	0,8336	0,8337	0,8392	0,8403	0,8194
(okno=12)	**std**	0,0804	0,0811	0,0827	0,0743	0,0818	0,0829	0,0792	0,0822	0,0881
rep_hydrofil2	**mean**	0,8288	0,8307	0,9906	0,8518	0,8339	0,8366	0,8389	0,9936	0,8279
(okno=12)(detrend)	**std**	0,0845	0,0829	0,0234	0,0787	0,0795	0,0861	0,0828	0,0191	0,0857
rep_hydrofil2	**mean**	0,8722	0,9966	0,8727	0,8663	0,8670	0,8691	0,8704	0,8774	0,8623
(okno=30)	**std**	0,0765	0,0139	0,0736	0,0756	0,0777	0,0798	0,0791	0,0753	0,0783
rep_hydrofil2	**mean**	0,8611	0,8659	0,9927	0,8411	0,8574	0,8619	0,8599	0,9932	0,8683
(okno=30)(detrend)	**std**	0,0737	0,0757	0,0195	0,0839	0,0780	0,0801	0,0819	0,0189	0,0716
rep_izoel	**mean**	0,8764	0,8633	0,9134	0,8771	0,8540	0,8551	0,9614	0,9166	0,8759
	std	0,0757	0,0803	0,0629	0,0777	0,0802	0,0822	0,0434	0,0652	0,0762
rep_izoel	**mean**	0,8593	0,8719	0,9930	0,8637	0,8548	0,8551	0,8446	0,9948	0,8619
(detrend)	**std**	0,0756	0,0805	0,0191	0,0774	0,0811	0,0781	0,0809	0,0166	0,0802
rep_izoel	**mean**	0,8668	0,8557	0,9181	0,8628	0,8502	0,8548	0,9620	0,9190	0,8739
(nekumulovane)	**std**	0,0810	0,0849	0,0631	0,0793	0,0862	0,0794	0,0457	0,0624	0,0793
rep_PHCurve	**mean**	0,9008	0,8912	0,9907	0,8936	0,8757	0,8749	0,8570	0,9940	0,9084
	std	0,0719	0,0723	0,0217	0,0707	0,0757	0,0771	0,0274	0,0176	0,0667
rep_PHCurve	**mean**	0,8673	0,8660	0,9930	0,8677	0,8709	0,8627	0,9234	0,9936	0,8703
(detrend)	**std**	0,0767	0,0800	0,0197	0,0775	0,0789	0,0755	0,0646	0,0185	0,0803

Tabulka A.12: Ohodnocení kvality reprezentací genu ND4.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,8941	0,8853	0,9087	0,8642	0,8514	0,8367	0,7893	0,9107	0,9040
	std	0,0704	0,0775	0,0656	0,0791	0,0799	0,0827	0,0870	0,0646	0,0644
rep_eiip	**mean**	0,8343	0,8412	0,9954	0,8298	0,8217	0,8336	0,8328	0,9952	0,8456
(detrend)	**std**	0,0832	0,0855	0,0157	0,0840	0,0882	0,0812	0,0825	0,0160	0,0795
rep_eiip	**mean**	0,7681	0,7853	0,7839	0,8762	0,7693	0,7642	0,7621	0,7930	0,7634
(nekumulovane)	**std**	0,0915	0,0917	0,0887	0,0772	0,0981	0,0982	0,0964	0,0871	0,0966
rep_disoc1	**mean**	0,9120	0,8852	0,9090	0,8773	0,8738	0,8587	0,9838	0,9060	0,9129
	std	0,0676	0,0736	0,0693	0,0759	0,0729	0,0802	0,0308	0,0722	0,0654
rep_disoc1	**mean**	0,8531	0,8517	0,9961	0,8454	0,8463	0,8438	0,8499	0,9951	0,8567
(detrend)	**std**	0,0809	0,0795	0,0146	0,0835	0,0815	0,0772	0,0815	0,0169	0,0820
rep_disoc1	**mean**	0,7602	0,7897	0,7763	0,8431	0,7684	0,7659	0,7554	0,7803	0,7619
(nekumulovane)	**std**	0,0932	0,0906	0,0870	0,0805	0,0929	0,0996	0,0932	0,0875	0,0933
rep_hydrofil2	**mean**	0,7749	0,7736	0,7870	0,8277	0,7766	0,7766	0,7669	0,7809	0,7627
(okno=4)	**std**	0,0964	0,0912	0,0917	0,0790	0,0943	0,0879	0,0971	0,0960	0,0921
rep_hydrofil2	**mean**	0,7714	0,7770	0,9937	0,8269	0,7682	0,7726	0,7712	0,9939	0,7613
(okno=4)(detrend)	**std**	0,0968	0,0922	0,0184	0,0834	0,0953	0,0917	0,0977	0,0178	0,0961
rep_hydrofil2	**mean**	0,7773	0,9972	0,7888	0,8134	0,7882	0,7821	0,7851	0,7964	0,7640
(okno=12)	**std**	0,0906	0,0126	0,0946	0,0860	0,0906	0,0899	0,0909	0,0877	0,0942
rep_hydrofil2	**mean**	0,7777	0,7784	0,9961	0,8138	0,7794	0,7774	0,7839	0,9957	0,7718
(okno=12)(detrend)	**std**	0,0956	0,0931	0,0142	0,0897	0,0939	0,0913	0,0899	0,0153	0,0930
rep_hydrofil2	**mean**	0,8184	0,9950	0,8169	0,8388	0,8147	0,8121	0,8188	0,8202	0,7871
(okno=30)	**std**	0,0843	0,0174	0,0835	0,0824	0,0846	0,0939	0,0883	0,0863	0,0917
rep_hydrofil2	**mean**	0,8116	0,7992	0,9961	0,8342	0,8130	0,8092	0,8188	0,9949	0,8052
(okno=30)(detrend)	**std**	0,0878	0,0898	0,0146	0,0874	0,0854	0,0834	0,0892	0,0165	0,0868
rep_izoel	**mean**	0,8996	0,8839	0,9002	0,8684	0,8490	0,8419	0,9824	0,9018	0,9016
	std	0,0662	0,0719	0,0655	0,0745	0,0797	0,0847	0,0301	0,0695	0,0706
rep_izoel	**mean**	0,8530	0,8572	0,9963	0,8277	0,8510	0,8467	0,8522	0,9951	0,8540
(detrend)	**std**	0,0784	0,0777	0,0138	0,0825	0,0788	0,0867	0,0815	0,0165	0,0828
rep_izoel	**mean**	0,8960	0,8860	0,9024	0,8743	0,8516	0,8499	0,9820	0,9034	0,9048
(nekumulovane)	**std**	0,0703	0,0665	0,0707	0,0736	0,0800	0,0819	0,0320	0,0653	0,0683
rep_PHCurve	**mean**	0,9128	0,8999	0,9829	0,8971	0,8994	0,8989	0,9850	0,9957	0,9197
	std	0,0684	0,0715	0,0282	0,0682	0,0699	0,0709	0,0284	0,0157	0,0612
rep_PHCurve	**mean**	0,8979	0,8959	0,9960	0,8867	0,8789	0,8862	0,9043	0,9958	0,8963
(detrend)	**std**	0,0697	0,0700	0,0148	0,0737	0,0751	0,0738	0,0638	0,0156	0,0678

Tabulka A.13: Ohodnocení kvality reprezentací genu ND4L.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,9111	0,9059	0,8810	0,8816	0,8570	0,8349	0,8419	0,8801	0,9106
	std	0,0655	0,0704	0,0706	0,0739	0,0828	0,0822	0,0847	0,0763	0,0671
rep_eiip	**mean**	0,8280	0,8344	0,9900	0,8407	0,8361	0,8341	0,8311	0,9908	0,8448
(detrend)	**std**	0,0826	0,0790	0,0230	0,0805	0,0785	0,0850	0,0876	0,0230	0,0793
rep_eiip	**mean**	0,7813	0,7992	0,7686	0,8661	0,7882	0,7796	0,7814	0,7648	0,7686
(nekumulovane)	**std**	0,0895	0,0830	0,0943	0,0775	0,0905	0,0929	0,0919	0,0942	0,0909
rep_disoc1	**mean**	0,8951	0,8881	0,8828	0,8759	0,8524	0,8359	0,9869	0,8836	0,9059
	std	0,0667	0,0750	0,0787	0,0736	0,0778	0,0878	0,0277	0,0729	0,0677
rep_disoc1	**mean**	0,8420	0,8396	0,9904	0,8356	0,8334	0,8291	0,8371	0,9889	0,8393
(detrend)	**std**	0,0862	0,0834	0,0224	0,0813	0,0853	0,0836	0,0845	0,0244	0,0844
rep_disoc1	**mean**	0,7847	0,8117	0,7598	0,8600	0,7984	0,7938	0,7882	0,7546	0,7802
(nekumulovane)	**std**	0,0913	0,0822	0,0964	0,0784	0,0863	0,0833	0,0886	0,0874	0,0938
rep_hydrofil2	**mean**	0,7908	0,7919	0,7733	0,8134	0,7844	0,7770	0,7824	0,7757	0,7646
(okno=4)	**std**	0,0909	0,0863	0,0889	0,0902	0,0900	0,0906	0,0867	0,0865	0,0935
rep_hydrofil2	**mean**	0,7801	0,7873	0,9897	0,8173	0,7778	0,7779	0,7862	0,9894	0,7774
(okno=4)(detrend)	**std**	0,0885	0,0859	0,0228	0,0931	0,0924	0,0867	0,0912	0,0235	0,0911
rep_hydrofil2	**mean**	0,8100	0,9948	0,8009	0,8287	0,8051	0,8099	0,8003	0,7989	0,7871
(okno=12)	**std**	0,0871	0,0177	0,0916	0,0852	0,0947	0,0918	0,0894	0,0863	0,0893
rep_hydrofil2	**mean**	0,7973	0,7957	0,9854	0,8146	0,7983	0,8009	0,8106	0,9864	0,7979
(okno=12)(detrend)	**std**	0,0931	0,0850	0,0278	0,0864	0,0893	0,0902	0,0936	0,0270	0,0866
rep_hydrofil2	**mean**	0,8392	0,9956	0,8664	0,8217	0,8332	0,8390	0,8362	0,8702	0,8422
(okno=30)	**std**	0,0872	0,0151	0,0747	0,0855	0,0783	0,0834	0,0811	0,0753	0,0826
rep_hydrofil2	**mean**	0,8314	0,8373	0,9879	0,8192	0,8234	0,8279	0,8411	0,9893	0,8412
(okno=30)(detrend)	**std**	0,0833	0,0835	0,0253	0,0833	0,0833	0,0798	0,0817	0,0243	0,0818
rep_izoel	**mean**	0,8746	0,8706	0,8790	0,8573	0,8563	0,8532	0,9860	0,8838	0,8749
	std	0,0743	0,0741	0,0754	0,0778	0,0794	0,0822	0,0284	0,0733	0,0771
rep_izoel	**mean**	0,8477	0,8562	0,9923	0,8357	0,8387	0,8454	0,8404	0,9914	0,8470
(detrend)	**std**	0,0825	0,0761	0,0204	0,0812	0,0838	0,0828	0,0772	0,0213	0,0830
rep_izoel	**mean**	0,8676	0,8702	0,8826	0,8584	0,8578	0,8474	0,9893	0,8862	0,8688
(nekumulovane)	**std**	0,0768	0,0771	0,0712	0,0739	0,0774	0,0777	0,0233	0,0736	0,0754
rep_PHCurve	**mean**	0,8947	0,8863	0,9821	0,8908	0,8876	0,8812	0,9897	0,9904	0,8981
	std	0,0707	0,0682	0,0325	0,0751	0,0708	0,0696	0,0241	0,0232	0,0697
rep_PHCurve	**mean**	0,8763	0,8710	0,9909	0,8713	0,8537	0,8594	0,9119	0,9922	0,8772
(detrend)	**std**	0,0751	0,0724	0,0220	0,0713	0,0806	0,0785	0,0645	0,0214	0,0740

Tabulka A.14: Ohodnocení kvality reprezentací genu ND5.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,9211	0,9064	0,9726	0,9081	0,8946	0,8802	0,8394	0,9761	0,9204
	std	0,0600	0,0710	0,0388	0,0611	0,0735	0,0758	0,0836	0,0356	0,0647
rep_eiip	**mean**	0,8770	0,8690	0,9950	0,8810	0,8802	0,8817	0,8808	0,9954	0,8796
(detrend)	**std**	0,0702	0,0783	0,0159	0,0740	0,0738	0,0744	0,0779	0,0160	0,0726
rep_eiip	**mean**	0,8373	0,8359	0,8621	0,8564	0,8361	0,8304	0,8328	0,8667	0,8316
(nekumulovane)	**std**	0,0804	0,0834	0,0758	0,0807	0,0818	0,0857	0,0855	0,0713	0,0837
rep_disoc1	**mean**	0,8964	0,8910	0,9708	0,8640	0,8651	0,8392	0,9317	0,9683	0,9027
	std	0,0681	0,0704	0,0402	0,0780	0,0769	0,0822	0,0593	0,0421	0,0662
rep_disoc1	**mean**	0,8462	0,8461	0,9966	0,8431	0,8322	0,8328	0,8368	0,9960	0,8396
(detrend)	**std**	0,0793	0,0795	0,0134	0,0814	0,0865	0,0829	0,0895	0,0148	0,0836
rep_disoc1	**mean**	0,8319	0,8454	0,8567	0,8731	0,8351	0,8354	0,8314	0,8671	0,8328
(nekumulovane)	**std**	0,0834	0,0826	0,0758	0,0714	0,0844	0,0841	0,0837	0,0731	0,0852
rep_hydrofil2	**mean**	0,8221	0,8214	0,8650	0,8546	0,8271	0,8222	0,8218	0,8592	0,8198
(okno=4)	**std**	0,0817	0,0832	0,0803	0,0728	0,0859	0,0874	0,0816	0,0709	0,0881
rep_hydrofil2	**mean**	0,8164	0,8296	0,9959	0,8453	0,8229	0,8279	0,8192	0,9958	0,8291
(okno=4)(detrend)	**std**	0,0835	0,0849	0,0146	0,0854	0,0876	0,0818	0,0845	0,0152	0,0868
rep_hydrofil2	**mean**	0,8041	0,9969	0,8560	0,8392	0,8097	0,8061	0,8124	0,8582	0,8130
(okno=12)	**std**	0,0847	0,0133	0,0763	0,0812	0,0855	0,0863	0,0878	0,0782	0,0805
rep_hydrofil2	**mean**	0,8149	0,8172	0,9964	0,8356	0,8067	0,8121	0,8166	0,9961	0,8124
(okno=12)(detrend)	**std**	0,0827	0,0820	0,0136	0,0876	0,0828	0,0849	0,0907	0,0146	0,0903
rep_hydrofil2	**mean**	0,8290	0,9957	0,8739	0,8590	0,8288	0,8260	0,8251	0,8752	0,8238
(okno=30)	**std**	0,0823	0,0149	0,0757	0,0777	0,0834	0,0909	0,0889	0,0735	0,0825
rep_hydrofil2	**mean**	0,8271	0,8279	0,9962	0,8467	0,8287	0,8274	0,8232	0,9951	0,8181
(okno=30)(detrend)	**std**	0,0826	0,0820	0,0140	0,0796	0,0833	0,0878	0,0865	0,0161	0,0874
rep_izoel	**mean**	0,9166	0,8982	0,9733	0,8858	0,8816	0,8568	0,9359	0,9728	0,9148
	std	0,0667	0,0696	0,0379	0,0733	0,0727	0,0783	0,0579	0,0389	0,0611
rep_izoel	**mean**	0,8683	0,8697	0,9968	0,8629	0,8561	0,8602	0,8656	0,9956	0,8698
(detrend)	**std**	0,0759	0,0775	0,0130	0,0807	0,0788	0,0784	0,0739	0,0151	0,0758
rep_izoel	**mean**	0,9136	0,9040	0,9752	0,8862	0,8831	0,8597	0,9440	0,9763	0,9217
(nekumulovane)	**std**	0,0651	0,0728	0,0374	0,0758	0,0737	0,0764	0,0513	0,0361	0,0673
rep_PHCurve	**mean**	0,9276	0,9142	0,9926	0,8968	0,9007	0,8973	0,9327	0,9961	0,9266
	std	0,0615	0,0624	0,0211	0,0676	0,0684	0,0757	0,0540	0,0146	0,0606
rep_PHCurve	**mean**	0,9051	0,8998	0,9960	0,8931	0,8936	0,8933	0,9348	0,9976	0,9057
(detrend)	**std**	0,0648	0,0689	0,0144	0,0730	0,0730	0,0688	0,0584	0,0119	0,0677

Tabulka A.15: Ohodnocení kvality reprezentací genu ND6.

Reprezentace	RF vzdálenost	euclidean	seuclidean	hamming	chebychev	cosine	correlation	spearman	jaccard	cityblock
rep_eiip	**mean**	0,9114	0,9023	0,9311	0,8922	0,8728	0,8527	0,8574	0,9253	0,9107
	std	0,0642	0,0668	0,0563	0,0708	0,0752	0,0758	0,0751	0,0598	0,0652
rep_eiip (detrend)	**mean**	0,8588	0,8493	0,9954	0,8622	0,8572	0,8603	0,8578	0,9941	0,8554
	std	0,0756	0,0746	0,0157	0,0803	0,0758	0,0794	0,0775	0,0178	0,0807
rep_eiip (nekumulovane)	**mean**	0,7902	0,8100	0,8326	0,8837	0,7763	0,7872	0,7853	0,8287	0,7916
	std	0,0870	0,0831	0,0805	0,0718	0,0879	0,0847	0,0927	0,0782	0,0836
rep_disoc1	**mean**	0,8873	0,8676	0,9287	0,8710	0,8564	0,8356	0,9790	0,9253	0,8933
	std	0,0716	0,0744	0,0599	0,0779	0,0799	0,0809	0,0339	0,0593	0,0699
rep_disoc1 (detrend)	**mean**	0,8464	0,8404	0,9936	0,8339	0,8413	0,8350	0,8381	0,9942	0,8443
	std	0,0813	0,0820	0,0182	0,0798	0,0813	0,0854	0,0825	0,0180	0,0811
rep_disoc1 (nekumulovane)	**mean**	0,7931	0,8307	0,8373	0,8661	0,7882	0,7863	0,7870	0,8332	0,7956
	std	0,0852	0,0753	0,0781	0,0746	0,0932	0,0884	0,0810	0,0807	0,0886
rep_hydrofil2 (okno=4)	**mean**	0,7834	0,7842	0,8328	0,8369	0,7879	0,7860	0,7811	0,8299	0,7798
	std	0,0780	0,0831	0,0881	0,0806	0,0862	0,0907	0,0823	0,0836	0,0881
rep_hydrofil2 (okno=4)(detrend)	**mean**	0,7783	0,7844	0,9938	0,8266	0,7867	0,7778	0,7877	0,9927	0,7769
	std	0,0897	0,0823	0,0182	0,0837	0,0869	0,0919	0,0868	0,0198	0,0843
rep_hydrofil2 (okno=12)	**mean**	0,7960	0,7984	0,8449	0,8284	0,7949	0,7966	0,8014	0,8510	0,7764
	std	0,0835	0,0850	0,0794	0,0809	0,0881	0,0832	0,0832	0,0786	0,0893
rep_hydrofil2 (okno=12)(detrend)	**mean**	0,7971	0,7991	0,9960	0,8247	0,7970	0,7994	0,7958	0,9953	0,7867
	std	0,0876	0,0850	0,0148	0,0784	0,0874	0,0887	0,0889	0,0158	0,0883
rep_hydrofil2 (okno=30)	**mean**	0,8311	0,9954	0,8937	0,8324	0,8340	0,8254	0,8260	0,8951	0,8316
	std	0,0800	0,0153	0,0720	0,0815	0,0845	0,0836	0,0831	0,0658	0,0780
rep_hydrofil2 (okno=30)(detrend)	**mean**	0,8340	0,8448	0,9944	0,8311	0,8319	0,8311	0,8381	0,9939	0,8406
	std	0,0842	0,0859	0,0171	0,0825	0,0789	0,0823	0,0812	0,0178	0,0770
rep_izoel	**mean**	0,8677	0,8438	0,9272	0,8507	0,8361	0,8298	0,9773	0,9282	0,8768
	std	0,0782	0,0759	0,0591	0,0831	0,0816	0,0806	0,0334	0,0572	0,0732
rep_izoel (detrend)	**mean**	0,8398	0,8474	0,9951	0,8252	0,8322	0,8357	0,8460	0,9949	0,8497
	std	0,0810	0,0764	0,0169	0,0859	0,0808	0,0832	0,0776	0,0165	0,0745
rep_izoel (nekumulovane)	**mean**	0,8801	0,8541	0,9323	0,8563	0,8382	0,8307	0,9763	0,9242	0,8662
	std	0,0719	0,0828	0,0584	0,0773	0,0828	0,0863	0,0367	0,0601	0,0753
rep_PHCurve	**mean**	0,8691	0,8640	0,9807	0,8602	0,8730	0,8761	0,9868	0,9957	0,8776
	std	0,0792	0,0765	0,0324	0,0788	0,0773	0,0766	0,0273	0,0157	0,0742
rep_PHCurve (detrend)	**mean**	0,8736	0,8756	0,9948	0,8560	0,8683	0,8757	0,9197	0,9949	0,8702
	std	0,0718	0,0723	0,0173	0,0780	0,0760	0,0761	0,0626	0,0168	0,0806

B. Obsah CD

Data adresář obsahující použitá data.

- **Fasta** adresář s fasta soubory všech genů.
- **data.mat** matlabovská datová skruktura obsahující všechny použité geny organismů.

Funkce adresář obsahující použité funkce.

- **rep_disoc1.m**
- **rep_eiip.m**
- **rep_hydrofil2.m**
- **rep_izoel.m**
- **rep_PHCurve.m**
- **Resample2.m**
- **RF_dist.m**
- **Test.m**

Barton_BP.pdf elektronická verze práce.

test.m skript použitý pro testování kvality reprezentací.

Literatura

[1] E. Kočárek. *Genetika*. Scientia, Praha, 2. vyd. edition, 2008.

[2] J. Flegr. *Evoluční biologie*. Academia, Praha, 2., opr. a rozš. vyd. edition, 2009.

[3] F. Cvrčková. *Úvod do praktické bioinformatiky*. Academia, Praha, vyd. 1. edition, 2006.

[4] Studijní materiály k předmětu bioinformatika. 2015. garant předmětu: I. provazník.

[5] I. Shmulevich ed. by: E. R. Dougherty. *Genomic signal processing and statistics*. Hindawi Publ. Corporation, New York, NY [u.a.], 2005.

[6] I. Cosic. Macromolecular bioactivity. *IEEE Transactions on Biomedical Engineering*, vol. 41(issue 12):1101–1114.

[7] V. Veljković, I. Ćosić, B. Dimitrijević, and D. Lalović. Is it possible to analyze dna and protein sequences by the methods of digital signal processing? *IEEE Transactions on Biomedical Engineering*, 32(5), 1985.

[8] Y. Liu, D. Li, K. Lu, Jiao, and Ping-An He. P-h curve, a graphical representation of protein sequences for similarities analysis. *Match*, (70.1: 451-466), 2013.

[9] W. Deng and Y. Luan. Dv-curve representation of protein sequences and its application. *Computational and Mathematical Methods in Medicine*, vol. 2014:1–8, 2014.

[10] Y., S. Yan, H. Xu, J. Han, X. Nan, P. He, and Q. Dai. Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation. *Evolutionary Bioinformatics*.

[11] Jia Wen and YuYan Zhang. A 2d graphical representation of protein sequence and its numerical characterization. *Chemical Physics Letters*, vol. 476(4-6):281–286, 2009.

[12] Hailong Hu. F-curve, a graphical representation of protein sequences for similarity analysis based on physicochemical properties of amino acids. *Match*, (73), 2015.

[13] Tingting Ma, Yuxin Liu, Qi Dai, Yuhua Yao, and Ping an He. A graphical representation of protein based on a novel iterated function system. *Physica A: Statistical Mechanics and its Applications*, vol. 403:21–28, 2014.

[14] Yan ping Zhang, Ya jun Sheng, Wei Zheng, Ping an He, and Ji shuo Ruan. Novel numerical characterization of protein sequences based on individual amino acid and its application. *BioMed Research International*, vol. 2015:1–8, 2015.

[15] Zhong Li, Geng, Pingan He, and Yao. A novel method of 3d graphical representation and similarity analysis for proteins. *Match*, (71), 2014.

[16] D. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, vol. 40(issue 9):1098–1101, 1952.

[17] Zhao-Hui Qi, Jun Feng, Xiao-Qin Qi, and Ling Li. Application of 2d graphic representation of protein sequence based on huffman tree method. *Computers in Biology and Medicine*, vol. 42(issue 5):556–563, 2012.

[18] H.J. Jeffrey. Chaos game rrepresentation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.

[19] M. Randic, J. Zupan, A.T. Balaban, D. Vikic-Topic, and D. Plavsic. Graphical representation of proteins. *Chemical Reviews*, 111(2):790–862, 2011.

[20] S. Basu, A. Pan, C. Dutta, and J. Das. Chaos game representation of proteins. *Journal of Molecular Graphics & Modelling*, 15(5):279–289, 1997.

[21] D.f. robinson a l.r. foulds. comparison of phylogenetic trees. doi: 10.1016/0025-5564(81)90043-2. isbn 10.1016/0025-5564(81)90043-2.

[22] Mathworks. statistics and machine learning toolbox user's guide [online]. 10.2. 2016, s. 4061-4066 [cit. 2016-05-01].

[23] Ncbi. ncbi common tree [online]. 2016 [cit. 2016-05-08]. dostupné z: http://www.ncbi.nlm.nih.gov/taxonomy/commontree/wwwcmt.cgi.

[24] Ch. d. michener and r. r. sokal. a quantitative approach to a problem in classification [online]. [cit. 2016-05-14]. doi: 10.2307/2406046. isbn 10.2307/2406046.

[25] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega.

[26] Ulrich Bodenhofer, Enrico Bonatesta, Christoph Horejs-Kainrath, and Sepp Hochreiter. msa: an r package for multiple sequence alignment. *Bioinformatics*, 31(24):3997–3999, 2015.

[27] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[28] Mathworks, (2016). bioinformatics toolbox: User's guide (r2016a). získáno z http://www.mathworks.com/help/pdf_doc/bioinfo/bioinfo_ug.pdf.