DIGITAL SIGNAL PROCESSING FOR GENOME CLASSIFICATION OF PLANTS

Robin Jugas

Bachelor Degree Programme (3), FEEC BUT

E-mail: xjugas00@stud.feec.vutbr.cz

Supervised by: Karel Sedlář

E-mail: sedlar@feec.vutbr.cz

Abstract: Numerical representations of DNA sequence underwent a significant development during last few years. Nevertheless, there are several techniques producing signals with various features. In this work, we present dendrogram construction for three different signal representations and its comparison with standard method based on multiple alignment of character sequences. Signal classification is performed as a cluster analysis using Euclidian metrics measuring pairwise distance between all pairs of signals aligned using dynamic time warping.

Keywords: classification, DNA signal, dtw

1. ÚVOD

Od roku 2001, kdy byl dokončen projekt sekvenace lidského genomu se bioinformatika dostala do popředí vědeckého zájmu. Standardním nosičem genetického kódu je znaková sekvence DNA, počítačové zpracování je však výpočetně náročné. Zde přichází ke slovu numerická sekvence DNA. Ta zachovává většinu informací nesenou bázemi a současně přináší mnoho výhod od rychlejšího zpracování po možnosti využití číslicového zpracování signálů. Podobně jako znakové sekvence mohou být i DNA signály zarovnány a pomocí shlukové analýzy klasifikovány. Právě toto je tématem článku – porovnat několik numerických reprezentací z hlediska klasifikace organismů.

2. PŘEVOD NA DNA SIGNÁL

DNA signál je číselná sekvence získaná ze znakové sekvence DNA dle metody numerické reprezentace, jichž je aktuálně široký výběr. Pro potřeby klasifikace genů jsem zvolil reprezentace kumulovanou a rozbalenou fází [1] a reprezentaci DNA walk [2], z důvodu jejich jednorozměrnosti, která se s výhodou uplatní při procesu zarovnání metodou DTW a výpočtu vzdáleností.

2.1. SIGNÁLOVÉ REPREZENTACE

Výchozí pro obě fázové reprezentace je komplexní reprezentace [1]. Nukleotidovým bázím (A, C, G, T) v sekvenci jsou přiřazeny komplexní čísla dle vztahu (1):

$$A = -1 + i \;\; C = -1 - i \;\; G = 1 + i \quad T = 1 - i \tag{1}$$

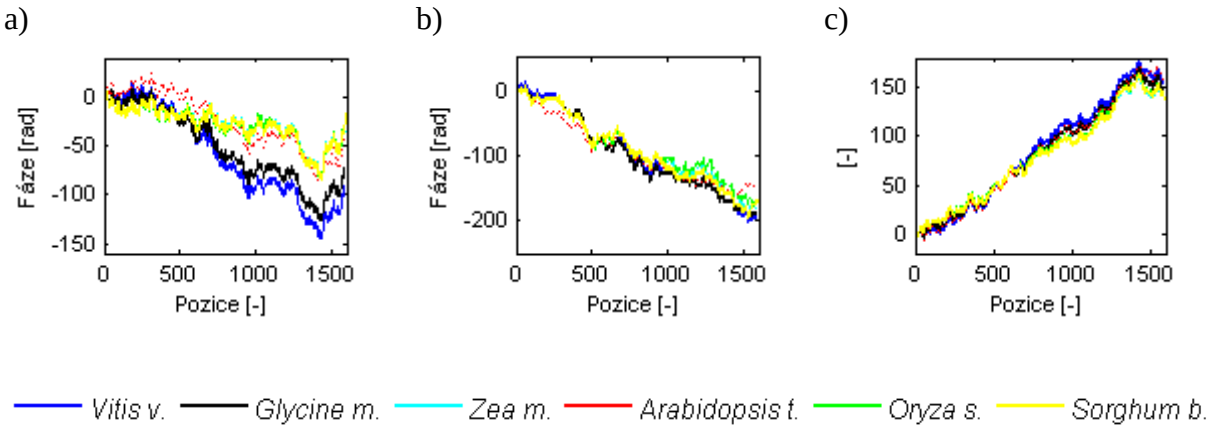
Kumulovaná fáze je kumulativním součtem všech fází komplexních čísel od počátku sekvence až po aktuální pozici. Rozbalená fáze je korigovanou fází elementů v sekvenci. Absolutní hodnota rozdílu fází dvou po sobě jdoucích je udržována na hodnotě menší než π.

DNA walk je reprezentace pro zobrazení DNA sekvence a pozorování korelací v širokém rozsahu [2]. V závislosti na typu báze v sekvenci *S* definujme přiřazení (2) na dané pozici *k*:

$$S[k] \in (C, G) \rightarrow x[k] = +1$$
$$S[k] \in (A, T) \rightarrow x[k] = -1 \quad (2)$$

2.2. MATERIÁL PRO ANALÝZU

Mitochondrie je buněčná organela, která obsahuje svoji vlastní specifickou sekvenci DNA, která je pro svůj evoluční význam častým zdrojem zkoumání. Pro analýzu jsem z mitochondriálních genů zvolil gen *cox1*, přítomný napříč bakteriálními a eukaryotickými druhy, a provedl analýzu pro šest náhodně vybraných rostlinných sekvencí získaných z databáze NCBI.



Obrázek 1: Ukázky reprezentací a) kumulovaná fáze b) rozbalená fáze c) DNAwalk

3. KLASIFIKACE ORGANISMŮ S VYUŽITÍM NUMERICKÝCH REPREZENTACÍ

Získané numerické sekvence byly zarovnány metodou dynamického borcení časové osy (DTW). Cílem této metody je vzájemná synchronizace sekvencí. Délka obou signálů je algoritmem přizpůsobitelná a signály mohou být i zkráceny. Z důvodu zachování podmínky omezení sklonu křivky u metody DTW [3], byly signály všech reprezentací nejdříve zbaveny lineárního trendu.

Mezi dvojicemi takto zarovnaných číselných sekvencí byla vypočítána Euklidova vzdálenost a metodou UPGMA sestrojen dendrogram každé reprezentace. Pro srovnání byl sestrojen opět dendrogram metodou UPGMA z původních vícenásobně zarovnaných znakových sekvencí na základě jejich proporcionálních vzdáleností. Výsledkem jsou tři fylogenetické stromy z numerických signálů a jeden znakový, který považujeme za referenční.

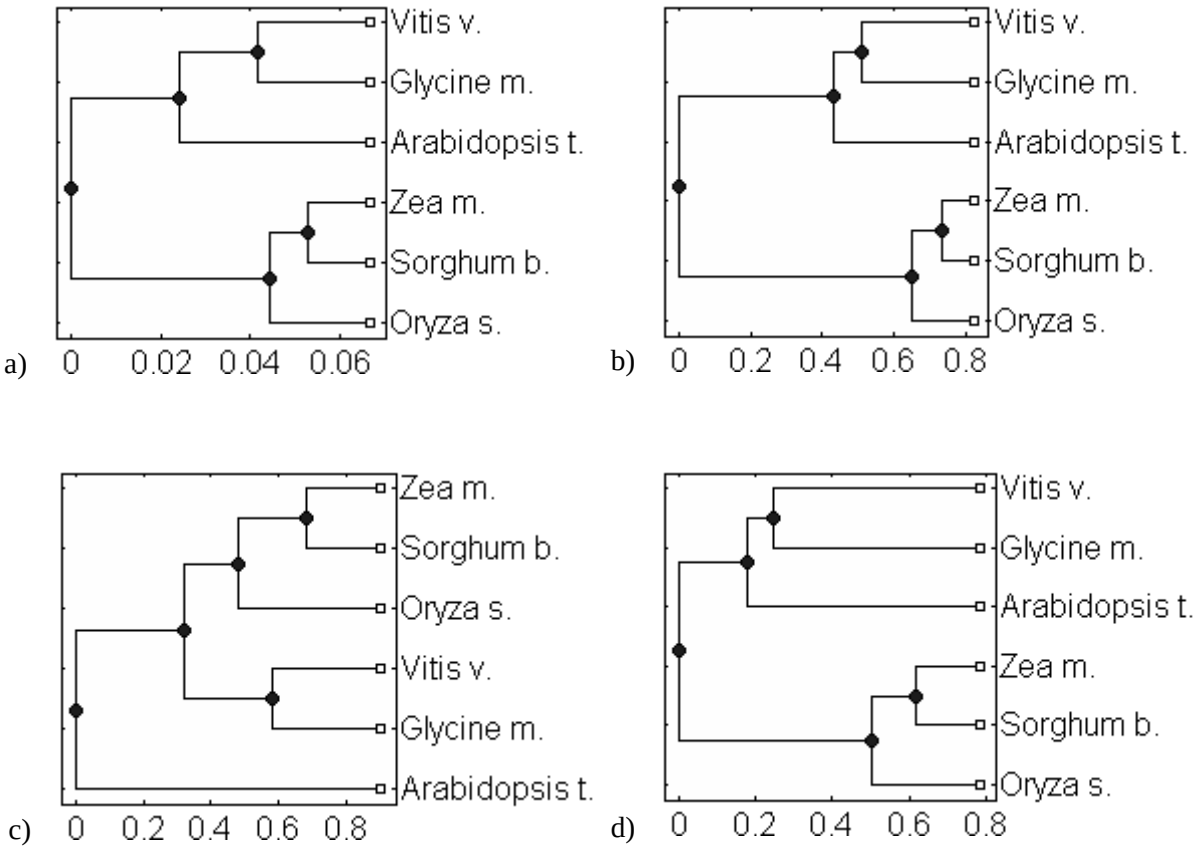
Pro srovnání fylogenetických stromů z jednotlivých reprezentací s výchozím znakovým stromem byly vypočítány korelační koeficienty mezi dvojicemi stromů na základě srovnání matic vzdáleností získaných z těchto stromů.

Kumulovaná fáze	Rozbalená fáze	DNA walk
0,9913	0,5800	0,9631

Tabulka 1: Korelační koeficienty znakového stromu a stromu z reprezentace

Největší míru korelace z Tabulka 1 ukazuje fylogenetický strom z reprezentace kumulované fáze, ale výsledek z reprezentace DNA walk je též velmi přesný. Všeobecně hodnoty nad 0.8 jsou považovány za přijatelné. To nesplňuje strom z reprezentace rozbalenou fází.

Z přehledu výsledných dendrogramů na Obrázek 2 jsou zjištěné korelace viditelné. Stromy kumulované fáze i DNA walk jsou oba správně taxonomicky roztříděny.



Obrázek 2: a) Znakové sekvence b) Kumulované fáze c) Rozbalené fáze d) DNA walk

4. ZÁVĚR

V článku je prezentováno srovnání numerických reprezentací z hlediska klasifikace organismů. Klasifikace pomocí numerických reprezentací při zarovnání metodou DTW podává stejně kvalitní výsledky jako při použití konvenčních znakových sekvencí DNA, je však méně časově náročná a dovoluje aplikaci dalších metod signálového zpracování. Záleží však na vhodné volbě numerické reprezentace, některé neposkytují pro klasifikaci dostatečně kvalitní výsledky.

REFERENCE

[1] P. D. Cristea, "Phase analysis of DNA genomic signals," *Proc. 2003 Int. Symp. Circuits Syst. 2003. ISCAS '03.*, vol. 5, pp. V–25–V–28, 2003.

[2] J. a Berger, S. K. Mitra, M. Carli, and A. Neri, "Visualization and analysis of DNA sequences using DNA walks," *J. Franklin Inst.*, vol. 341, no. 1–2, pp. 37–53, Jan. 2004.

[3] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust.*, vol. 26, no. 1, pp. 43–49, Feb. 1978.