CLASSIFICATION OF PROKARYOTIC ORGANISMS BASED ON COMPRESSED WHOLE GENOME SIGNALS

Karel Sedlář

Master Degree Programme (2), FEEC BUT

E-mail: xsedla74@stud.feec.vutbr.cz

Supervised by: Helena Škutková

E-mail: skutkova@feec.vutbr.cz

Abstract: Modern classification of organisms is based on molecular data. These methods rely on multiple alignment of sequences of characters which make them computationally demanding. Only small parts of genomes can be compared in reasonable time. In this paper, the conversion of the whole genome sequences to cumulative phase signals is presented. Dyadic wavelet transform is used for lossy compression of signals by redundant frequency bands elimination. Signal classification is then performed as a cluster analysis using Euclidian metrics where multiple alignment is replaced by dynamic time warping.

Keywords: cumulative phase, whole genome, wavelet transform, dynamic time warping

1. ÚVOD

Klasifikace organismů je jednou ze základních otázek biologie. Jelikož hlavním nositelem dědičnosti je DNA, je porovnávání organismů založeno na molekulárních znacích. Přitom nové techniky sekvenace umožňují levné sestavení celého genomu jednotlivých organismů, zvláště pak prokaryotických, u kterých je genom tvořen jediným kruhových chromozomem. Klasické metody komparace jsou ale založené na vícenásobném zarovnání znakových sekvencí, které je výpočetně velmi náročné, pro více sekvencí s délkou nad 100 kbp prakticky nemožné. Klasifikace se tak provádí na úrovní genů (stovky až jednotky tisíc bp), které ale nemusí dobře popisovat vývoj celého organismu, jen vývoj tohoto konkrétního genu. Při nesprávné volbě genu je pak celá klasifikace chybná. Převodem sekvence znaků na signál kumulované fáze zjistíme, že takový signál je zčásti redundantní a dokáže si uchovat význačné charakteristiky i po masivní ztrátové kompresi. Znaková sekvence tuto vlastnost nemá. Komprimované signály pak umožní porovnávat celé chromozomy nebo i genomy organismů. Přitom jako obdobu zarovnání sekvencí je vhodné pro signály použít dynamické borcení časové osy (dynamic time warping, DTW [3]).

2. GENOMICKÝ SIGNÁL

Prvním krokem klasifikace je převod sekvenčních dat do podoby číselné řady - signálu. Signály jsou pak na rozdíl od sekvencí pro účely klasifikace komprimovatelné.

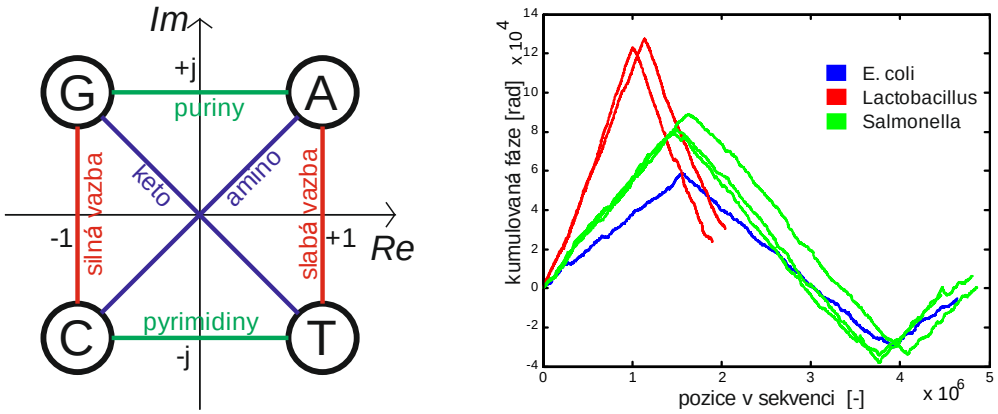
2.1. KONVERZE SEKVENCE NA SIGNÁL

Numerických reprezentací DNA existuje celá řada, přičemž pro naše účely jsme zvolili metodu kumulované fáze [1]. Každý z nukleotidů A, C, G, T vyskytujících se v DNA je promítnut do komplexní roviny tak, že přiřazením vhodného komplexního čísla zůstanou vhodně zachovány informace o jejich chemické podobnosti, viz. Obrázek 1. Hodnotu kumulované fáze každé pozice v sekvenci lze pak určit ze vzorce:

$$\theta_C = \frac{\pi}{4}[3(n_G - n_C) + (n_A - n_T)] \quad (1)$$

kde n_A, n_C,n_G, n_T jsou četnosti výskytu jednotlivých nukleotidů do aktuální pozice sekvence.

Na následujícím obrázku vpravo je 6 genomických signálů z vybraných organismů.



Obrázek 1: Komplexní reprezentace nukleotidů (vlevo) a signály kumulované fáze (vpravo).

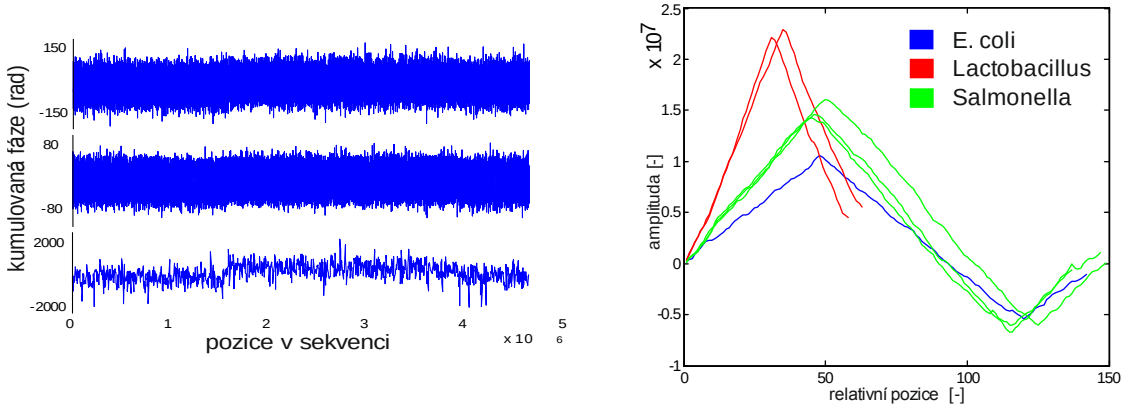
2.2. KOMPRESE SIGNÁLŮ

Komprese signálů byla provedena s použitím dyadické vlnkové transformace [2]. Ta se od Fourierovy transformace liší použitím jiné bázové funkce, neboli vlnky:

$$\psi_{\lambda,\vartheta}(t) = \frac{1}{\sqrt{\lambda}}\psi\left(\frac{t-\vartheta}{\lambda}\right) \tag{2}$$

kde $\psi(t)$ je mateřská vlnka, λ dilatace (stlačení či roztažení) mateřské vlnky a ϑ časové posunutí vlnky. Pro dyadickou vlnkovou transformaci pak platí $\lambda=2^m$, kde $m>0$.

Vybrané signály byly rozloženy se stupněm rozkladu 15, tedy celkem na 16 frekvenčních pásem. Pro účely rozkladu lze s výhodou použít Haarovy vlnky. Jedná se o tvarově nejjednodušší obdélníkovou vlnku, díky níž je výpočet transformace velmi rychlý. Jak je patrné z Obrázku 1, délka signálů je obvykle v řádu miliónů vzorků a více. Rozkladem (Obrázek 2:) jsme dokázali, že typický tvar signálu je nesen nejnižším pásmem. Ostatní pásma s vyššími frekvencemi nesou informaci o jednotlivých nukleotidech a kratších úsecích signálu. Tyto informace přitom pro klasifikaci nejsou podstatné. Nejnižší pásmo pak dokáže celý signál reprezentovat pouze pomocí vzorků v počtu o několik řádů nižším než před kompresí. Celkově bylo dosaženo při zachování hlavních charakteristik signálů kompresního poměru 0,01% [4]. Ukázku 3 vysokofrekvenčních pásem pro *E. coli* a soubor komprimovaných signálů je možné vidět na Obrázku 2. Amplitudy vysokofrekvenčních pásem jsou vůči amplitudě celého signálu o několik řádů nižší a celkový tvar signálu tak ovlivňují pouze nepatrně. Pro komprimovaný signál už jsou obě osy bezrozměrné.

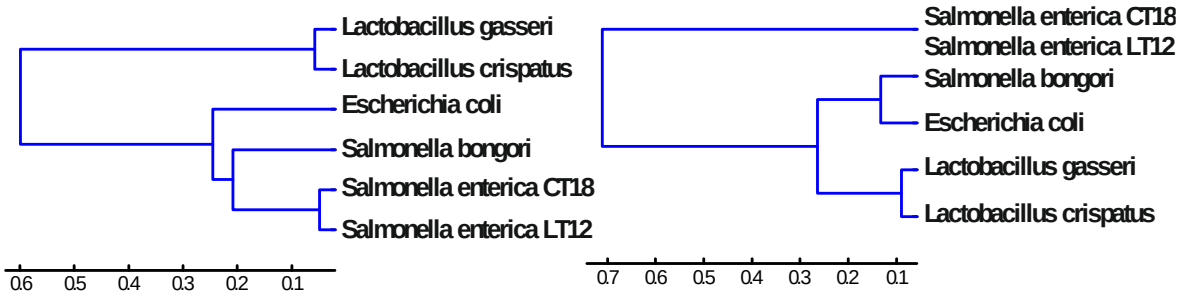


Obrázek 2: Vysokofrekvenční pásma po rozkladu (vlevo) a komprimované signály (vpravo).

3. KLASIFIKACE SEKVENCÍ

Z uvedených příkladů jsou patrné rozdíly v délce sekvencí. Před samotným porovnáním dvou sekvencí je potřeba takovou dvojici vždy zarovnat. Zarovnání nesmí být náhodné, signály nelze libovolně zkracovat. Pro zarovnání bylo použito dynamické borcení časové osy (DTW) [3], které optimálně zarovnává dva signály při minimalizaci jejich vzájemné vzdálenosti. Některé vzorky signálu mohou být duplikovány nebo vypuštěny. DTW je tak velmi podobné zarovnání dvou znakových sekvencí Needleman-Wunschovým algoritmem. Podobný je i jeho výpočet probíhající na základě tabulky lokálních vzdáleností. Vzdálenosti všech dvojic sekvencí jsme pak definovali jako jejich euklidovskou vzdálenost. Jelikož je tato vzdálenost bezrozměrná, byla asociační matice signálů normalizována a následně vyobrazena v dendrogramu průměrných vzdáleností.

Metoda celogenomové klasifikace je srovnána s dendrogramem sestrojeným ze znakových dat 16S rRNA, což je v současnosti nejpoužívanější metoda klasifikace prokaryot. Tyto sekvence mají typicky kolem 1500 bp a jsou definované pro jednotlivé druhy. Tato metoda tak například nerozliší dva poddruhy salmonely. Porovnáním celého chromozomu ale tyto dva organismy rozlišit dokážeme. Původní metoda navíc chybně klasifikuje rod salmonely, kdy druh *bongori* pokládá za bližší bakterii *E. coli*.



Obrázek 3: Dendrogramy klasifikovaných signálů (vlevo) a znakových sekvencí (vpravo).

4. ZÁVĚR

Současná klasifikace prokaryotických organismů je založena na porovnávání pouze krátkých úseků genomu, i když často máme k dispozici kompletní sekvenci celého kruhového chromozomu. Celogenomová komparace je tak limitována pouze výpočetní náročností algoritmů pro porovnávání genomických sekvencí. Námi představený algoritmus využívá převodu genomické sekvence na genomický signál. Tento signál je dále se ztrátou masivně komprimován pomocí dyadické vlnkové transformace, avšak při zachování svých charakteristických rysů potřebných pro porovnání dvou signálů. Celková výpočetní náročnost je tak výrazně nižší než u znakových metod, neboť kompresí podvzorkované signály jsou kratší. Jak je dokumentováno na posledním obrázku, tato nová metoda navíc odstraňuje problémy, které mohou nastat při klasifikaci pouze na základě krátkých úseků sekvencí.

REFERENCE

[1] Cristea, P. D.: Conversion of nucleotides sequences into genomic signals. Journal of Cellular and Molecular Medicine, 2002, 6(2), 279–303

[2] Daubechies, I.: Ten lectures on wavelets, CBMS-NSF conference series in applied mathematics. SIAM Ed, 1992

[3] Myers, C.S. and Rabiner, L.R.: A comparative study of several dynamic time-warping algorithms for connected word recognition. The Bell System Technical Journal 60:7, 1981, 1389–1409

[4] Salomon, D.: Data Compression: The Complete Reference. London: Springer Science+Business Media, LLC, 2007, ISBN 1-84628-602-6