BIOINFORMATICKÝ NÁSTROJ PRO KLASIFIKACI BAKTERIÍ DO TAXONOMICKÝCH KATEGORIÍ NA ZÁKLADĚ SEKVENCE GENU 16S RRNA

Autor: Nikola Valešová
Vedoucí: Ing. Stanislav Smatana

Motivace

Po dlouhou dobu bylo možné bakterie analyzovat pouze jejich kultivací. Mnohé druhy bakterií jsou však nekultivovatelné, a proto nebyly vůbec zjistitelné. Díky nedávnému pokroku v sekvenování s vysokou průchodností je nyní možné účinně vyšetřovat mikrobiální komunity a analyzovat druhy bakterií, které se v nich nacházejí. Pomocí sekvenování jsme schopni získat různé sekvence 16S rRNA bakterií ve vzorku z prostředí, avšak pro detekování, o jaké organismy se jedná, je následně nutno tyto sekvence klasifikovat, což je složitý problém.

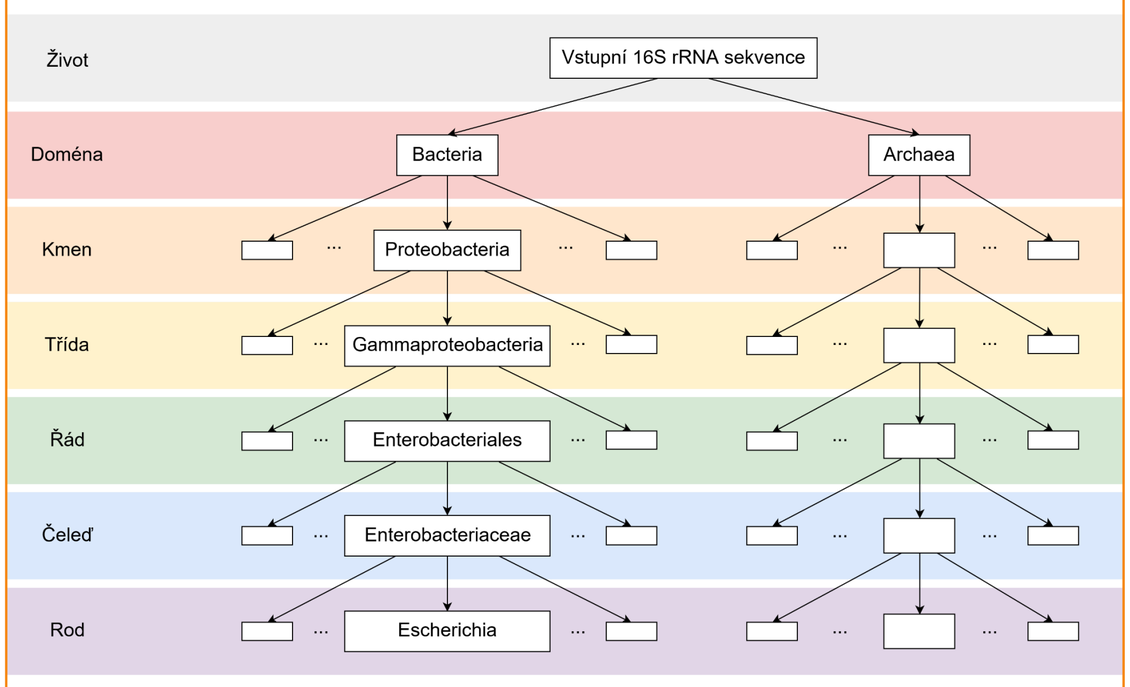
Princip metody klasifikace

Navržená metoda klasifikace bakterií je založena na stromové struktuře taxonomických kategorií. Celý klasifikátor se skládá ze stromu parciálních klasifikátorů s topologií respektující taxonomický strom. Klasifikace vstupního vzorku začíná v horním klasifikátoru rozlišujícím mezi bakteriemi a *archaea* a vstupní posloupnost sestupuje stromem podle klasifikovaných označení. Parciální klasifikátory představují známé metody strojového učení (např. SVM, *decision tree* a *nearest centroid*) a jejich cílem je klasifikovat dané bakterie a přiřadit jim třídu na nižší taxonomické úrovni.

Tato metoda klasifikuje sekvence na základě jejich 16S rRNA genu. 16S rRNA se nachází v genomu všech druhů bakterií a lze v ní najít dva typy oblastí − konzervované části a variabilní úseky, které procházejí rychlými genetickými změnami, a proto jsou vhodné pro odlišení druhů. Každá sekvence 16S rRNA je dlouhý řetězec složený ze čtyř nukleotidů − A, C, G a T. Sekvence 16S rRNA mohou navíc obsahovat mutace, inserce a delece, proto mohou mít různé sekvence různou délku. Algoritmy strojového učení vyžadují, aby jejich vstupní vektory měly stejné rozměry, proto se ze vstupní sekvence extrahuje k-merové spektrum, které se použije pro klasifikaci. Díky k-merovému spektru je možné transformovat každou 16S rRNA sekvenci do numerického vektoru, kde každá hodnota představuje počet výskytů odpovídajícího dílčího řetězce z původním sekvence.

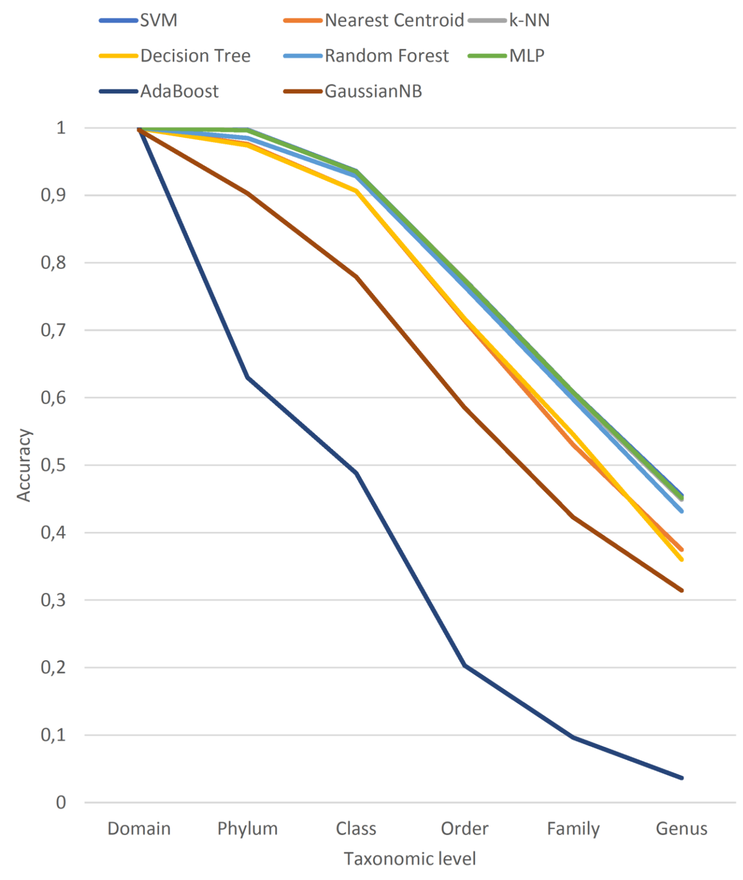
Pro minimalizaci chyby klasifikace je prezentovaná metoda založena na stromové struktuře taxonomického stromu. Celý klasifikátor je rozložen do více parciálních klasifikátorů na všech úrovních taxonomie od domény až po rod. Tímto způsobem by mělo být možné dosáhnout vyšší přesnosti predikce, protože každý klasifikátor rozlišuje pouze mezi několika málo třídami na nižší úrovni taxonomie.

Stromová struktura klasifikátorů



Obrázek 1: Stromová struktura klasifikátorů vytvořená podle hierarchie taxonomických tříd. Každý obdélník představuje jeden paricální klasifikátor a jeho označení ukazuje klasifikaci vstupní sekvence na odpovídající úrovni taxonomie. Je uveden příklad klasifikace bakterie *Escherichia coli*. Parciální klasifikátory jsou spojeny šipkami, které označují pořadí sekvenční klasifikace skrze strom klasifikátorů.

Dosažené výsledky



Obrázek 2: Závislost přesnosti klasifikace na taxonomické úrovni pro jednotlivé typy dílčích klasifikátorů. Pro vyhodnocení byl použit dataset BLAST 16S. Přesnost byla vyhodnocena cross-validací leave-one-out. Pro získání výsledků byl použit k-mer 5, protože se ukázalo, že je časově a paměťově efektivní a zároveň poskytuje dobrou přesnost.