

自主課題研究概要

R を用いた哺乳類の生物種分類

電子情報学類 3 年 301 安藤 雄気

1. 研究の動機と背景

遺伝子配列を用いて生物種の分類ができるかについて興味があり、このテーマを調べることにした。

2. 関連研究と本研究

関連研究に吉田正和氏らが JHES2008 で発表した研究がある。これは rRNA の配列を用いて生物種の分類、予測をするというものであり、本研究も基本的にこれに同じだが、データが膨大になるため、哺乳類に限って調べることにした。

3. 実験データと実験方法

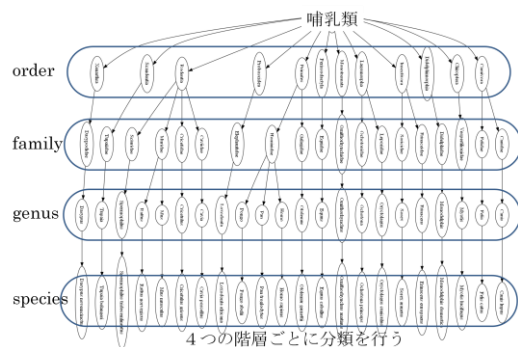


図1：対象とした哺乳類の生物種分類

本研究で対象としたのは22種の哺乳類に関する209個のrRNAデータである。図1のように目、科、属、種の4つの階層ごとに別々の分類問題として学習、予測を行った。特徴情報としてはRNA配列の5-gramの頻度、1024次元とした。分類器はRのkernlabパッケージに入っているksvm関数を使い、精度は交差検定(leave-one-out cross-validation)で求めた。

4. 実験結果

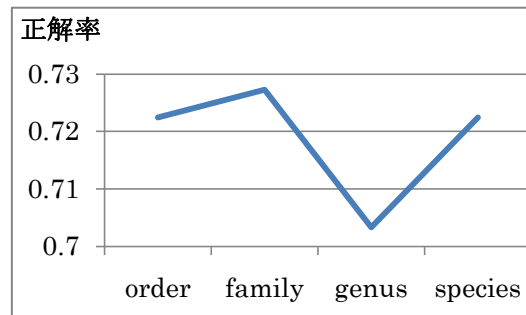


図2：各階層の分類の正解率

図2のように各階層の分類の正解率はおおむね70%となった。本来は階層が小さくなるにつれて正解率は減少するはずだが、ここではgenus(属)が特に低くなっている。また、生物種ごとの分類の正解しやすさにおいてはヒトやチンパンジーなどデータ量の多い物の方がイヌやネコなどデータ量の少ないものに比べ、正解が多いことが分かった。

5. 考察

図2のように減少具合がおかしくなった原因は用いたデータ量が少なかったことにあると思われる。また、データ量の多い生物種の方が正解の多かったことはサポートベクターマシンの性質をよくあらわしていると言える。今回の研究ではおおむね予想通りの結果が得られ、遺伝子配列から生物種が分類できることも分かった。とても興味があったので、次はクラスタリング等他の方法を用いて進化の関係を調べてみたいと思う。