Propostas para o trabalho final

Proposta A

Atualmente o GenBank representa um poderoso banco de dados onde as sequências já publicadas em trabalhos científicos são depositadas. Entretanto, o site só permite baixar um arquivo com os dados por grupo taxonômico contendo todos os marcadores (genes) disponíveis, ou precisamos selecionar gene a gene para termos arquivos diferentes para o grupo taxonômico por gene. Minha ideia é fazer uma função que em que a entrada será um arquivo com tudo o que está disponível no GenBank para meu grupo taxonômico de interesse (o arquivo será adquirido através do site <http://www.ncbi.nlm.nih.gov/genbank/>) , e irá gerar vários arquivos de texto, cada um contendo a lista de táxons para cada marcador (gene). Além disso, o arquivo que é baixado do GenBank apresenta um cabeçalho grande antes de iniciar a sequência propriamente dita. Pretendo, nesta mesma função, inserir um argumento que me permita remover todo esse cabeçalho e deixar apenas as informações que são relevantes, como o nome da espécie e o número de acesso do GenBank. O *default* do argumento seria FALSE, mantendo todo o cabeçalho; e se TRUE, as informações consideradas excedentes seriam removidas.

Proposta B

O site do GenBank não atualiza os nomes dos táxons das sequências depositadas de acordo com as revisões taxonômicas feitas. Com isso, preciso sempre verificar se a nomenclatura do táxon do GenBank continua a mesma. Minha ideia é fazer uma função que me permita associar os nomes dos táxons das sequências disponíveis no GenBank ao histórico taxonômico da espécie disponível no site “Amphibian Species of the World” (<http://research.amnh.org/vz/herpetology/amphibia/>). Assim, poderei de maneira otimizada atualizar a nomenclatura de todos os táxons disponíveis no meu arquivo.