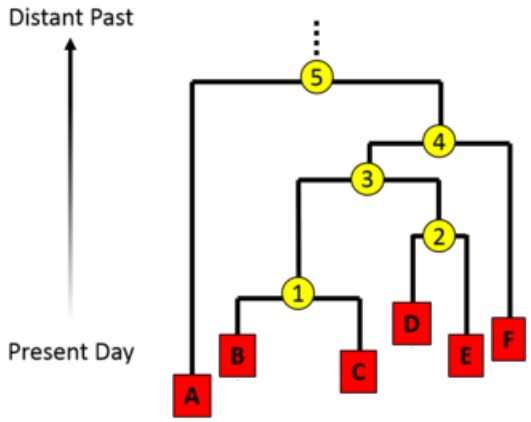
Rekonstrukce mezer v ancestrálních sekvencích

Ancestrální rekonstrukce je metoda určená pro odvození sekvencí genů, proteinů a organizmů, které se v současné době už s velkou pravděpodobností v přírodě nevyskytují, můžeme však nalézt jejich příbuzné sekvence, které se z těch ancestrálních v průběhu milionů let evoluce vyvinuly. Pomocí ancestrální rekonstrukce se tedy můžeme dopátrat, jak vypadal nejbližší společný předek vybraných organizmů.

Ancestrální strom je (pro naše potřeby binární) fylogenetický strom, kde listové uzly tohoto stromu jsou v dnešní době existující sekvence, zatímco vnitřní uzly stromy představují jejich společné předky.



Za posledních padesát let vznikla řada algoritmů, které jsou schopné vypočítat nejpravděpdobnější podobu sekvencí ve vnitřních (žlutých) uzlech na základě sekvencí na listech stromu. Mezi nejpoužívanější přístupy patří pravděpodobnostní metoda maximum-likelihood, která stejně jako ostatní vychází z vícenásobného zarovnání sekvencí.

Velkým problémem soudobých algoritmů je skutečnost, že k výpočtu pravděpodobností používají tzv. evoluční matice, zachycující pravděpodobnost záměny jedné aminokyseliny (nukleotidu) za druhou v průběhu evoluce a nejsou schopny pracovat s inzercí/delecí, reprezentovanou ve vícenásobném zarovnání symbolem pro mezeru (-). Tato skutečnost vede k bobtnání ancestrálních sekvencí, jelikož vícenásobné zarovnání bývá výrazně delší než jednotlivé sekvence v něm obsažené a maximum-likelihood se snaží dosadit symbol (ne mezeru) na každou pozici v zarovnání. Ancestrální sekvence, sestavené ze sekvencí o délce cca 300 aminokyselin, tak často nabývají i více jak dvojnásobné délky.

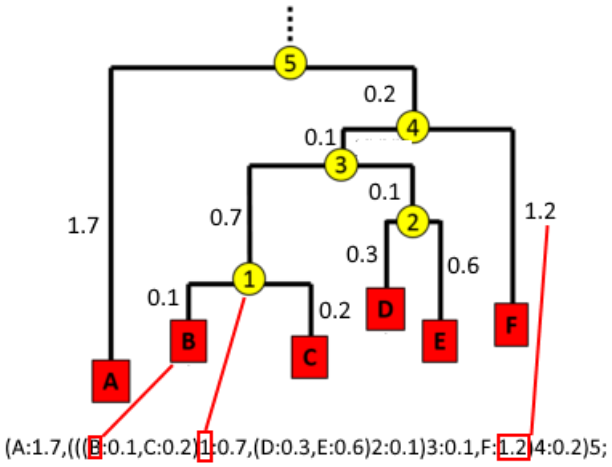
Z tohoto důvodu je nezbytné do ancestrálních sekvencí mezery tzv. zrekonstruovat, tedy nahradit některé ze znaků v ancestrální sekvenci za symbol mezery. Z toho vyplývá i zadání pro tento projekt.

Potřebné soubory

V souborech k projektu najdete kromě tohoto zadání další tři soubory:

msa.fasta – soubor s vícenásobným zarovnáním existujících sekvencí (sekvence na listových uzlech stromu). Sekvence jsou zaznamenány ve FASTA formátu.

tree.tre – soubor obsahující fylogenetický strom ve formátu Newick (viz níže).



ancestrals.csv – obsahuje tzv. posterior probabilities, tedy pravděpodobnost, s jakou se v dané ancestrální sekvenci a na dané pozici vyskutují jednotlivé aminokyseliny. První sloupec tabulky označuje vnitřní uzel, ke kterému se řádek vztahuje. Ve druhém sloupci je uvedeno číslo sloupce ve vícenásobném zarovnání (počítáno od 1). Zbylé sloupce pak udávají pravděpodobnosti pro jednotlivé aminokyseliny.

Zadání

Vypracujte script v jazyku Python, který provede následující:

1) Načtěte vícenásobné zarovnání ze souboru msa.fasta
2) Načtěte strukturu fylogenetického stromu ze souboru tree.tre. K tomuto účelu můžete napsat vlastní parser nebo použít některou ze stávajících knihoven, sloužících k práci s Newick formátem. Jednou z takových knihoven je knihovna Phylo, která je součástí balíčku Biopython. (pozn. V přiloženém souboru tree.tre jsou názvy vnitřních uzlů uloženy na pozici vyhrazené pro bootstrap hodnoty a s využitím knihovny Phylo je k nim tedy možné přistoupit s využitím atributu confidence)
3) Načtěte data ze souboru ancestrals.csv a na jejich základě odhadněte nejvíce pravděpdobné ancestrální sekvence, včetně doplnění ancestrálních mezer.

Tipy:

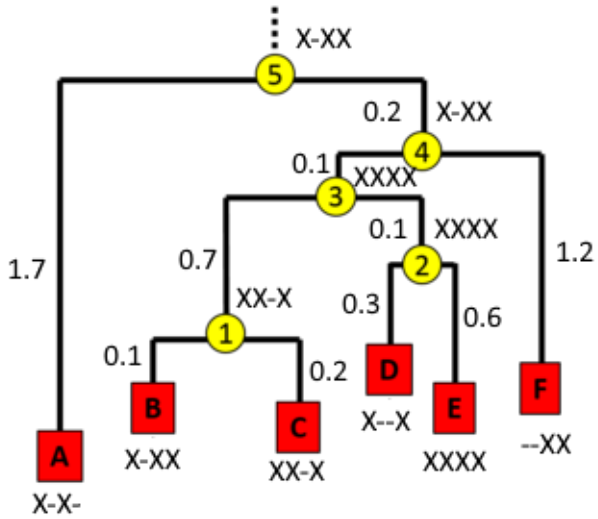
Základní podobu ancestrální sekvence pro daný uzel (bez doplnění ancestrálních mezer) můžete získat jednoduše tak, že pro každý ancestrální uzel a pro každou jeho pozici vyberete vždy tu nejvíce pravděpodobnou aminokyselinu z daného řádku v souboru ancestrals.csv.

Ancestrální mezery je nezbytné doplnit se zvážením struktury fylogenetického stromu. Pokud se tedy podíváme na obrázek uvedený výše, ancestrální mezery pro vnitřní uzel 3 by měly být určeny na základě listů B,C,D a E.

Nejjednodušším řešením, jak určit přítomnost ancestrální mezery pro uzel 3 na pozici 10 je tedy nahlédnout do souboru msa.fasta a podívat se na desátý sloupec u sekvencí B,C,D a E. Pokud počet mezer převažuje nad jinými znaky, je možné na tuto pozici doplnit mezeru i v ancestrálním uzlu.

Toto řešení zanedbává velké množství faktorů a z pohledu evoluce není přesné, jelikož ne všechny listy ve stromu jsou od ancestrálního uzlu stejně vzdálené (všimněte si různé délky větví na přiloženém obrázku) a některé sekvence by tudíž měly mít v rozhodování vyšší váhu než jiné. Pro získání plného počtu bodů by jste tedy měli při rozhodnutí uvážit nejenom frekvenci mezer, ale také vzdálenost, která dělí listy stromu od vnitřního uzlu.

Pokud bychom uvážili ancestrální uzel 1 a k němu náležející sekvence B a C a za předpokladu, že délka evoluční větve mezi uzlem 1 a sekvencí B je 0.1 a v sekvenci B se nachází mezera, a mezi 1 a C je délka 0.2 a mezera se zde nenachází, pak v ancestrální sekvenci by se mezera neměla vyskytnout (viz obrázek níže). Pro uzly nacházející se ve stromu výše by se pak délky větví sčítaly. Cesta z B do 3 je tedy 0,1+0,7.



Požadované výstupy:

Python script s kódem řešícím daný problém.

Složka se soubory obsahující vyřešené ancestrální sekvence. Každá sekvence by měla být vložena do separátního souboru, označeného jako node_X.fas, kde X je číslo ancestrálního uzlu. Uvnitř souboru by měla být uvedena vyřešená ancestrální sekvence o délce odpovídající délce sekvencí v MSA. Na každé pozici by měl být vložen buďto znak vybraný na základě „posterior probabilities“ nebo symbol pro mezeru „-“. (příklad: MQR-T-MG--ALI)