Predikce vlivu mutací na funkci proteinu

V praxi existuje velké množství výpočetních nástrojů pro predikci vlivu aminokyselinových mutací na proteiny. Může se jednat o specializované nástroje, určené pro predikci vlivu mutací na stabilitu, solubilitu, aktivitu, atd. Stejně tak ale můžeme použít i obecné nástroje, snažící se identifikovat pouze to, zda mutace bude neutrální (pozitivní) nebo škodlivá.

Úkolem tohoto projektu bude jeden takový prediktor sestavit. Zatímco mnohé nástroje se vydávají cestou strojového učení nad různě rozsáhlými a kvalitními datasety, v tomto projektu bude vaším úkolem sestavit knowledge-based prediktor, využívající evoluční informaci a sadu fyzikálně chemických vlastností.

Soubory:

tree.tre – soubor s fylogenetickým stromem v newick formátu

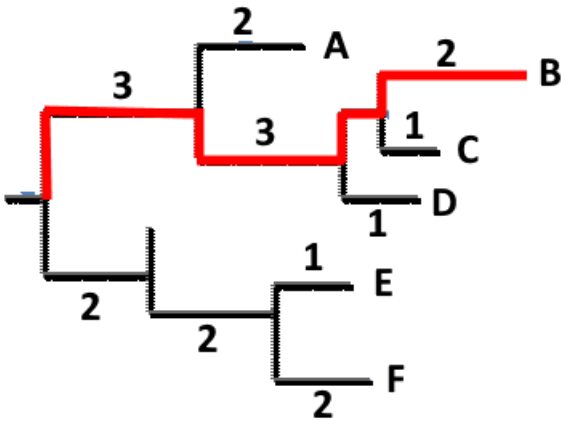
msa.fasta – soubor s vícenásobným zarovnáním sekvencí

aaindex.txt – soubor s maticemi fyzikálně chemických vlastností

Postup:

Skóre pro mutaci na dané pozici ve vícenásobném zarovnání by mělo být určeno jako kombinace evolučních vzdáleností, skóre konzervovanosti a rozdílu fyzikálně-chemických vlastností mezi původní a mutovanou aminokyselinou.

1) Vypočítejte evoluční vzdálenost pro každou listovou sekvenci ve fylogenetickém stromu. Tato evoluční vzdálenost je udána jako celková vzdálenost od listu ke kořeni stromu. Pro listovou sekvenci B na přiloženém obrázku bude evoluční vzdálenost rovna hodnotě 8.



2) Proveďte min-max normalizaci evolučních vzdáleností všech listů tak, aby nejdelší cesta ve stromu měla hodnotu 1 a všechny délky se tudíž pohybovaly v rozmezí <0,1>. Takto normalizované hodnoty použijte jako váhy pro sekvence vícenásobného zarovnání v následujícím kroku.
3) Vypočtěte skóre konzervovanosti pro každou aminokyselinu v každém sloupci vícenásobného zarovnání. Ve výsledku by jste měli mít dvacet hodnot (jednu pro každou standardní

aminokyselinu) pro každý sloupec vícenásobného zarovnání. Tato hodnota by měla být určena jako váhovaný podíl výskytu dané aminokyseliny. Pokud tedy budeme počítat skóre konzervovanosti pro aminokyselinu A, pak hodnota pro daný sloupec bude:

$$\frac{\sum_{i=0}^{n}(A_i * w_i)}{\sum_{i=0}^{n} w_i},$$

kde i značí řádek vícenásobného zarování, A_i má hodnotu 1, pokud se aminokyselina A vyskytuje na daném řádku zarovnání (jinak 0), a w_i značí váhu sekvence na daném řádku vypočítanou v předešlém kroku.

4) Sestavte si matici hodnot fyzikálně chemických vlastností ze souboru aaindex. Uvedené formátování je identické s tím, jaké můžete najít v databázi aaindex. Pokud se podíváte na první sloupeček, pak hodnota 0.61 je hodnotou pro aminokyselinu A a 1.53 je hodnotou pro aminokyselinu L. Obdobně v dalších sloupcích. Každou ze tří fyzikálně chemických vlastností opět min-max normalizujte, aby jste získali hodnoty v rozmezí <0,1>.

A/L	R/K	N/M	D/F	C/P	Q/S	E/T	G/W	H/Y	I/V
0.61	0.60	0.06	0.46	1.07	0.	0.47	0.07	0.61	2.22
1.53	1.15	1.18	2.02	1.95	0.05	0.05	2.65	1.88	1.32

5) V posledním kroku pronásobte váhovanou konzervovanost získanou v kroku tři s tabulkou fyzikálně chemických vlastností a určete rozdíl mezi původní a mutovanou aminokyselinou.

Příklad: Původní aminokyselina je A. Skóre konzervovanosti pro A je 0.25. Fyzikálně chemické parametry (po normalizaci) pro aminokyselinu A jsou 0.2, 0.4 a 0.7. Mutujeme na aminokyselinu C, která má skóre konzervovanosti 0.35. Fyzikálně chemické parametry pro C jsou 0.3, 0.5 a 0.1.

Výsledné skóre rozdílu A a C je (0.25*0.2 - 0.35*0.3) + (0.25*0.4 – 0.35*0.5) + (0.25*0.7 – 0.35*0.1)

Poznámka: Za původní aminokyselinu považujte tu, která se v daném sloupečku nachází v prvním řádku vícenásobného zarovnání (první sekvenci zarovnání tedy berte jako svou query sekvenci – je takto i označena). Pokud se v query sekvenci na dané pozici vyskytuje znak mezery (pomlčka), vložte pomlčku rovněž do celého sloupečku na výstupu.

Výstup:

Jako výstup je požadován komentovaný Python script a .csv soubor s výsledky v následujícím formátu:

Soubor bude mít celkem 21 řádků. Prvním řádkem bude hlavička, kde první sloupeček bude označen jako AA a další budou čísla sloupečků vícenásobného zarovnání, počítáno od jedničky.

Následujících dvacet řádků pak budou hodnoty pro jednotlivé aminokyseliny pro každý sloupec zarovnání. Aminokyseliny budou na řádcích v pořadí, odpovídajícím pořadí v seznamu aaindex, tedy ARNDCQEGHILKMFPSTWYV.