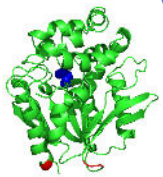


Využití variačních autoenkodérů pro ancestrální rekonstrukci sekvencí

autor: Ing. Pavel Kohout
vedoucí: Ing. Miloš Musil, Ph.D.

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Motivace

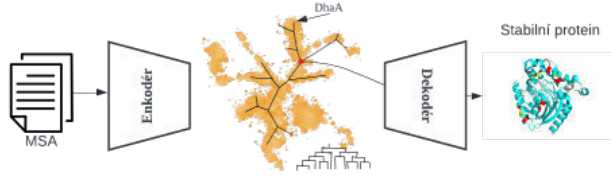
Proteiny jsou základními stavebními prvky živých organismů. Každý protein je tvořen unikátní sekvencí aminokyselinových jednotek, které jednoznačně určují jeho prostorové uspořádání, biologickou funkci a vlastnosti. Proteinové sekvence se skládají z desítek až tisíců aminokyselin, přičemž v přírodě se vyskytuje 20 variant standardních aminokyselin. To vytváří obrovský kombinatoriální prostor, který není možné systematicky prohledat současnými metodami. Pouze malá část tohoto prostoru je však funkční a vyskytuje se v přírodě. Právě proto jsou pro úspěšnou aplikaci proteinů v průmyslu obvykle využívány upravené přírodní varianty. Jednou z kýžených vlastností je stabilita. Stabilita je nejčastější limitací proteinů pro jejich využití v **biomedicínských** nebo **biotechnologických** aplikacích. Proto je existence metody, která je schopna generovat sekvence blízké těm přírodním avšak s vylepšenými vlastnostmi jako je stabilita, velmi žádaná.

Ancestrální rekonstrukce sekvencí

Rekonstrukce evolučních předchůdců (ASR), tzv. ancestrálních sekvencí, určitého proteinu se ukázala jako účinná metoda pro tvorbu stabilnějších a aktivnějších proteinů. Tato metoda zkoumá evoluční vztahy mezi zarovnanými existujícími proteiny (MSA) a za pomoci fylogenetických stromů vytváří jejich evoluční předchůdce, které kýžené vylepšené vlastnosti často vykazují.

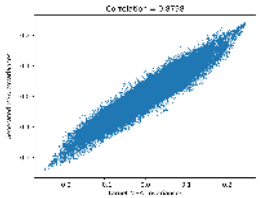
Náš přístup

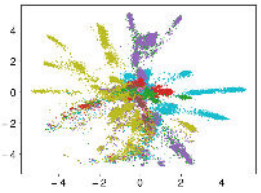
Jedním ze slibných směrů, který je schopen využívat pokročilé matematické modely společně s obrovským množstvím sekvenčních dat, je aplikace strojového učení v oblasti proteinové optimalizace. Zejména latentní modely (variační autoenkodéry (VAE)) ukázaly slibné vlastnosti, jelikož jejich struktura dokáže zachytit evoluční závislosti, obdobně jako fylogenetické stromy, tím že mapuje ancestrální sekvence blíže do centra latentního prostoru. Proto by se VAE mohly stát alternativou ke konvenčním, výpočetně náročným metodám využívajícím fylogenetické stromy pro ancestrální rekonstrukci sekvencí. Jako test vhodnosti využití latentních modelů jako alternativní metody ASR jsem provedl případovou studii VAE na proteinové rodině haloalkan dehalogenáz.



Evaluace modelu

Generování proteinových sekvencí je složitý proces kvůli dlouhým a drahým laboratorním experimentům. Proto byla sledována schopnost modelu generovat sekvence statisticky podobné těm ve vstupním datasetu společně se strukturními vlastnostmi vytvořeného latentního prostoru.





Evoluční strategie

- *Simulace řízené evoluce*: náhodné mutace sekvence → mapování do prostoru →ohodnocení, další kolo
- *Metoda přímé evoluce*: byly rekonstruovány body podél trajektorie do centra latentního prostoru
- *Strategie adaptace kovariační matice*: multikriteriální optimalizace využívající evoluční algoritmus a *pareto* frontu

Výsledky

Výstupy evolučních strategií byly podrobně statisticky vyhodnoceny a na základě vzniklých profilů byly vybrány varianty pro **laboratorní** experimenty. Ty prokázaly, že navržená metoda je schopna vytvořit funkční varianty s velkým počtem substitucí (až 45).