Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Michal Filippi
Název práce	Predikce sekundární struktury proteinu pomocí hlubokých neuronových sítí
Rok odevzdání	2017
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	David Hoksza **Role** Vedoucí
Pracoviště	KSI MFF UK

Text posudku:

Cílem předložené diplomvé práce bylo studium stávajících přístupů pro predikci sekundární struktury proteinů s důrazem na nejnovější přístupy využívajících hluboké neuronové sítě (DNN), implementaci vlastního DNN přístupu a jeho vyhodnocení vzhledem ke state-of-the-art metodám. Toto téma je aktuální jak vzhledem k důležitosti samotné úlohy predikce sekundární struktury proteinů, kdy predikční algoritmy jsou stále daleko od přesné predikce, tak s ohledem na informatický framework použitý pro řešení tohoto problémy, tj. hluboké neuronové sítě. Jednalo se tedy o hluboce interdisciplinární práci, která vyžadovala jak důkladné pochopení principů sekvenční a strukturní proteomiky, tak strojového učení.

V první části práce student představuje základní koncepty nutné k pochopení problému, tj. stavební prvky proteinů a jednotlivé úrovně struktury proteinu. Následuje část popisující problém predikce sekundární struktury a metodiky a metriky vyhodnocování predikcí, které závisí na typech sekundární struktury, které jsou predikovány. Následuje popis hlubokých neuronových sítí a to v kontextu nástrojů pro predikci sekundární struktury. Tato část tedy zároveň představuje přehled metodologie DNN a průřez state-of-the art nástrojů pro predikci sekundární struktury od metod používaných od 80. let 20. století až po současnost.

Výzkumná část pak popisuje jednotlivé přístupy, které byly v rámci práce implemetnovány a testovány. Jedná se o použití LSTM paměťových buněk, modifikace kroku využití evoluční informace formou pozičně specifických skórovacích matic (PSSM), zahrnutí zjednodušené predikce 3D struktury jako vlastností pro zlepšení učení sekundární struktury a úpravu metod pro vyhlazování predikce.

Všechny uvedené přístupy byly testovány a porovnány s nejlepšími publikovanými predikčními metodami. Student ukázal, že je možné vylepšit predikční možnosti těchto přístupů pro některé typy sekundárních struktur, zjednodušit používané modely (což má vliv na čas nutný pro naučení sítě), nebo zrychlit učení pomocí speciálního přístupu k PSSM. Naopak se ukázalo, že použití zjednodušené terciární struktury překvapivě nepřináší zvýšení predikčních schopností. Nicméně bylo ověřeno, že kdyby byly použity vlastnosti odvozené ze správné struktury, pak se schopnost predikce výrazně zvýší (tj. metodologie použití struktury je správná, ale použitá aproximace byla přílišná).

Student pracoval samostatně a průběžně, což bylo patrno z pravidelných konzultací. Je myslím vhodné vyzdvyhnout spektrum problémů, kterým se student musel věnovat. Nešlo jenom o metody DNN, kdy bylo třeba do důsledků pochopit a implementovat všechny

uvažované přístupy (jenom implementace stávajících řešení, jejichž architektura nebyla plně specifikována bylo náročná, nemluvě o nutnosti implementace na paralelní infrastruktuře) ale i o nutnost proniknout do čistě bioinformatických metod multiple sequence alignmentu, strukturní bioinformatiky, nastudovat datové formáty a přístupy, které se v daných doménách používají a to vše zakomoponovat do společného frameworku.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 28. August 2017 **Podpis**