OPONENTSKÝ POSUDEK DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení:	**Adámek**	Jméno:	**Petr**	Osobní číslo:	**368212**
Fakulta:	**Fakulta biomedicínského inženýrství**				
Studijní program:	**Biomedicínská a klinická technika**				
Studijní obor:	**Přístroje a metody pro biomedicínu**				
Název práce:	**Predikce a vizualizace proteinových sekundárních struktur na genomových datech**				

II. HODNOCENÍ DIPLOMOVÉ PRÁCE

	Kritéria hodnocení práce	Počet bodů
1.	Splnění cíle a vhodnost struktury obsahu diplomové práce z hlediska zadaného tématu (splnění zadání). (0 – 30)* Komentář: každé zadání, resp. každá část či věta ze zadání musí mít jasný odraz ve zpracované práci!, pouze zcela splněné zadání může být ohodnoceno max. 20 body. Podle rozsahu části v zadání, která není zcela vhodně či úplně zpracována, se snižuje ekvivalentně hodnota 20 bodů. Uvedení cíle v úvodu práce je povinné, a pokud není uvedeno, student přichází o 10 bodů. 30 celkových bodů může obdržet naprosto bezchybná a velmi precizně zpracovaná práce (to ale není standardní situace, spíše mimořádná).	20
2.	Teoretická úroveň a využití dostupné literatury v diplomové práci. (0 – 30)* Komentář: zde je velmi důležitá úloha oponenta a to následující: pokud je většina textu převzata, pak student získává max. 5 bodů, pokud je vše psáno slovy studenta, pak může získat max. 15 bodů, k tomu je možné připočítat max. 15 bodů za vhodné a ucelené zpracování dostupných pramenů, tj. je uveden současný stav v samostatné kapitole (5 bodů), významné relevantní zdroje jsou komentovány včetně popisu výběru (strategie výběru) těchto zdrojů (5 bodů) a použité zdroje jsou všechny a vhodně citovány, je posuzováno také složení citovaných zdrojů, tj. aktuálnost a vztah k tématu, obecné publikace jako matematické vzorce apod. se nepočítají do plnohodnotných citací, lze vypočítat poměr takovýchto citací, tj. užitečné/neužitečné a velikost tohoto poměru je třeba promítnout do bodování (5 bodů).	15
3.	Formální náležitosti a úprava obsahu diplomové práce (úroveň psaní, označení struktury textu, grafy, tabulky, citace v textu, seznam použité literatury apod.). (0 – 10)* Komentář: v současné době mají studenti k dispozici jak literaturu s popisem jak zpracovat odborný text na PC, mají znalosti a dovednosti a není tudíž třeba brát ohled na nedostatky z hlediska zpracování na PC, takže se předpokládá, že práce má obsah tvořen desetinným tříděním, zde lze hodnotit i orientaci v práci včetně odkazů mezi jednotlivými typy položek v textu včetně číslování rovnic, obrázků, tabulek a grafů (1 bod), práce obsahuje důležité položky z hlediska typu práce (2 body), kvalita obrázků (1 bod), množství překlepů (1 bod za nepatrné množství), v práci by se měla objevovat pouze standardní odborná terminologie a to zejména v českém jazyce (je třeba hodnotit schopnost vyjadřovat se technickým jazykem – 2 body), grafy jsou tvořeny podle zásad (viz tolerance a vliv statistického zpracování – 1 bod), u grafů a tabulek jsou patřičné legendy a vše je čitelné (1 bod), jsou dodržena citační pravidla podle ISO690 a ISO690-2 (1 bod).	5
4.	Rozsah realizačních prací (SW, HW), aplikovaných vědomostí a znalostí, úroveň metodologického zpracování a závěrů práce. (0 – 30)* Komentář: pokud je práce kombinací teoretických odvození (4 body – lze nahradit publikací v AJ), modelování a simulace (4 body), SW implementace (4 body) a též technické realizace (4 body – lze nahradit patentem či užitným vzorem) a 4 body ještě za komplexní funkčnost a to jak SW, tak i HW výstupu, pak může získat až 20 bodů. Pokud práce obsahuje správnou strukturu včetně diskuse výsledků (5 bodů – min. 2 strany A4) a závěrů (5 bodů – min. 1 strana A4), pak může být připočteno dalších 10 bodů. Celkem tedy 30 bodů za velmi komplexní a bezchybnou práci včetně uplatnění výsledků práce v rámci projektů, publikací, patentů či užitných vzorů.	10
5.	**Celkový počet bodů**	50

* Slovní hodnocení uveďte v komentáři.

České vysoké učení technické v Praze, Fakulta biomedicínského inženýrství
Katedra přírodovědných oborů, nám. Sítná 3105, 272 01 Kladno
tel.: +420 224 359 915, www.fbmi.cvut.cz

III. NÁVRH OTÁZEK K OBHAJOBĚ

1. Jaké jsou celkové statistiky metody pro všechny známé organismy? Tj. jaké jsou hodnoty precisssion/recall přes organismy, pro které jsou lokace BST2 známy?

2. Rodina BST je uvedena v PFAM jako PF16716 i s přislušným HMM. Jak si stojí prosté použití tohoto HMM ve srovnání s metodou použitou v této práci? Proč vlastně nebyla tato a podobné možnosti (tj. vyhledávání čistě na základě sekvence) diskutována?

3.

IV. CELKOVÉ HODNOCENÍ ÚROVNĚ VYPRACOVÁNÍ DIPLOMOVÉ PRÁCE

Hodnocení**:	A (výborně)	B (velmi dobře)	C (dobře)	D (uspokojivě)	E (dostatečně)	F (nedostatečně)
Počet bodů:	100 - 90	89 - 80	79 - 70	69 - 60	59 - 50	< 50
	❑	❑	❑	❑	X	❑

** v případě hodnocení F (nedostatečně) uveďte podrobný komentář

Diplomovou práci hodnotím výše uvedeným klasifikačním stupněm a ~~doporučuji~~/nedoporučuji k obhajobě.

V. KOMENTÁŘ

Práce je relativně komplexní ve smyslu výčtu nástrojů, které je možné použít pro predikci, což vidím jako plus práce. Dále na základě vyhodnocení na několika organismech se zdá, že skript vede na detekci příslušných genů. Jak dobře je ovšem těžko posoudit, protože chybí srovnání např. s (libovolně naivním) sekvenčním přístupem. Následuje seznam komentářů vztahující se k různým aspektům práce, včetně formátování práce, řešeršní části, popisu softwaru a softwaru samotného.

- Práce obsahuje množství překlepů a gramatických chyb, např.

1. "vpřípadě" -> "v případě"
2. "aminokyselinovými zbytky a o molekulární váze mezi" -> "a" je navíc
3. "U eukaryot je situace od dost" -> o dost
4. "nebývá o moc delší, nežli při oblast" -> "pri" navic
--- "Je dostupný na: bioinformatics.ysu.edu/tools/OrfPredictor.html" -> tecka navic
--- ... výše uvedené je po stranu 18, dále neuvádím.

- Některé tvrzení jsou zvláštní, např.
---- "kód programu byl vytvořen v programovacím jazyku PERL, což je vhodné zejména pro uživatele operačního systému Windows," - PERL zcela jistě není specifikum Windows
---- "Reprezentace genetického kódu zde využívá tři proměnné, seznamy typické pro Python 3." - Co znamená, že seznamy jsou "typické" pro Python?

- Formátování práce je velice chabé:
--- Není nikde používána kurziva, ačkoli na mnoha místech by její použití zvýšilo přehlednost
--- Seznamy jsou někdy uvedeny jako odstavce: např. str 20, kde je 10 odstaveců, kdy i-ty odstavec začíná vždy PRAVIDLO i:
--- Odstavce často obsahují pouze jednu větu a nedávají tudíž příliš smysl
--- Odstavce jsou někdy odděleny prázdnou řádkou, jindy ne.

- Rešeršní část je nekonzistentní, popisy různých nástrojů působí spíše tak, že autor náhodně vytáhl popisy z článků a nesnažil se je zpracovat jako nějaký homogenní popis, který by umožnil rozumné srovnání. Důsledkem je, že o různých nástrojích se dozvíme různé charakteristiky, např. u některých je popsán formát vstupu, jinde výstupu,

někde je detailně popsáno jak funguje jejich algoritmus, jinde toto není uvedeno vůbec. Někde se dozvíme, na jakých datech byl machine learning přístup trénován, jinde toto opět chybí.

- Popis softwarové části by jistě prospěl pseudokód, který by umožnil jednodušší porozumění jednotlivým částem popisu kódu.

- Vzhledem k faktu, že výstupem práce má být softwarový nástroj, moje nejváženější připomínky směřují k sofwarové části diplomové práce:

1. Standardem je publikovat softwarový kód ve formě repozitáře (např. GitHub), aby bylo možné kód stáhnout, podle instrukcí instalovat a spustit. Repozitář pak typicky obsahuje tedy i instrukce o možnostech programu a "manuál".
2. Kód skriptu je jeden ohromný soubor, který ani vnitřně není strukturovaný, tj. v principu jde o jednu obrovskou funkci. Výsledek je, že kód bude obtížně udržovatelný a kýmkoli rozšiřitelný (což tedy asi ani není vplánu s ohledem na neexistenci repozitáře, kde by komunita měla k výstupům práce přístup)
3. Adresářová struktura projektu taky není jasně strukturovaná. Tj. není jasné, kde jsou testovací data, co jsou pouze testovací výstupy (které by ani neměly být nutné, protože by měly být jednoduše vygenerovatelné)
4. Program nemá žádné parametry a změna vstupu se musí řešit změnou příslušné řádky v kódu, což mi přijde opravdu šílené a rozhodně jde proti větě v zadání, která říká, že "Výstupem práce bude nástroj pro predikci sekvence proteinu Bst2 a jiných příbuzných protivirových genů, případně identifikace Bst2 genů v nových organizmech". S nemožností parametrizovat skript to půjde obtížně.
5. Nejsou dobře ošetřené vstupy, např. na začátku se skript zeptá na kód ošetření ambiguitního vstupu. Když se předá prázdný vstup, program spadne na "ValueError: invalid literal for int() with base 10: ''
6. Protože autor používá Python, bylo by jednoduché použít virtuální prostředí, které by definovalo všechny dependence. Takto je extrémně obtížné, protože je třeba tyto dependence s jejich verzemi dohledat v práci a pak je jednu po druhé instalovat, nebo pouštět program, čekat kdy spadne na "ModuleNotFoundError" a doinstalovat dependenci (s tím, že pak nebude dependence nutně ve správné verzi)
7. Důsledek neexistence virtuálního prostředí je, že výsledky nejsou reprodukovatelné, resp. nejsou jendoduše reprodukovatelné, protože je složité získat přesně stejné verze knihoven, které používal autor. Stejně pak puštění formou modifikace kódu, místo jednoduché parametrizace skriptu značně komplikuje práci. Skripty k získání dat pro tabulky v sekci 5. 6. myslím v dodaných datech ani nejsou.
8. Výsledný HTML soubor není validní HTML, jak lze ověřit v libovolném online validátoru a tudíž může být problém se zobrazením v některém z prohlížečů. Validní HTML by měl být o to menší problém, že se jedná v podstatě jenom o základní HTML s obrázky.
9. Kód není dokumentovaný a obsahuje mnoho zakomentovaných částí. To bych u projektu, který je odevzdán jako diplomoví prcáce nečekal.

S ohledem na to, že cílem diplomové práce je softwarový projekt, k němuž mám silné výhrady, tak si nejsem jist jestli, ačkoli bodové hodnocení je 50 bodů a tedy dostatečně, je diplomová práce obhajitelná.

Jméno a příjmení: RNDr. David Hoksza, Ph.D.
Organizace: MFF UK Katedra softwarového inženýrství
Kontaktní adresa: Malostranské nám. 2/25, 118 00 Praha 1

Podpis: .....

Datum: .....