Oponentský posudek na diplomovou práci Víta Dohnálka

Vít Dohnálek sepsal diplomovou práci s názvem Evoluce importu proteinů do mitochondrií. Už z názvu vyplývá, že jde o práci velmi ambiciózní, která je však rozsahem spíše střídmá (naštěstí), obsahově však přesto velmi bohatá. Název práce je po mém soudu trochu zavádějící, protože práce se z velké části věnuje evoluci dvou mitochondriálních proteinů Tom40 a především Sam50. Až ve finální části pak popisuje stručně i evoluci dalších proteinů, které se podílejí na importu proteinů do mitochondrií.
Pokud jsem práci pochopil správně, tak hlavním cílem bylo zjistit, zda *Giardia intestinalis* skutečně neobsahuje protein Sam50.
Práce na úvod přináší literární přehled studované problematiky, který je velmi čtivý a vhodně doplněný obrázky. Autora bych však rád upozornil, že nevhodně používá termín β-list jak jako ekvivalent anglického β-sheet, tak častěji β-strand. β-list je však český překlad β-sheet - β-strand se překládá jako β-hřeben. Na závěr literárního přehledu je i kapitola o metodách predikce 3D struktur proteinů. Asi bych tam tuto kapitolu vlastně ani nevyžadoval, ale pokud tam je, tak z mého pohledu zasloužila větší pozornost – je založená na velmi málo publikacích spíše staršího data a neposkytuje úplně přesný přehled o metodách predikce 3D struktur. Speciálně u *ab-initio* predikcí se změnily jak používané přístupy, tak i přesnost metod.
Pokud se však v úvodu vyskytuje kapitola o predikci 3D struktur, čekal bych tam i kapitolu o metodách detekce homologů, protože ta je pro výsledky práce neméně důležitá.
Práce má na straně 14 velmi stručně definované čtyři cíle, které by si, myslím, zasloužily lepší vysvětlení i propojení s kapitolami výsledků a diskuse, protože některé z výsledků práce se mi těžko spojovali s uvedenými cíli (např. experimentální data k Tom40 bez β-signálu).
Kapitola Metodika má dvě části – stručnou, která popisuje stěžejní část autorovy práce (tedy tu bioinformatickou) a dlouhou, která popisuje laboratorní metody. Nemyslím si, že je délka vždy důležitá, ale v tomto případě se domnívám, že by si bioinformatická část zasloužila podrobnější metodiku. U prakticky žádného s použitých nástrojů není uvedeno v jakém nastavení byl použit. U MSA se píše že byly použity dva nástroje, ale není uvedeno kdy byl použit který a proč. Z Modelleru byly použity některé skripty, ale není jasné které. Podstatná část metodiky je pak až v části výsledky, ale ani tam ne vždy tak, aby bylo možné podle metodiky postupovat. Úplně mi chyběla část, která by se asi v klasické parazitologické práci jmenovala Materiál. Odkud pocházejí sekvence? Ve výsledkové části jsou na několika místech zmíněny různé uniprot respektive swissprot datasety, ale které verze datasetů to jsou? Dnes nejspíš budou tyto datasety větší – myslím, že by bylo správné, aby součástí takové práce byla i přiložená data, kde budou jednotlivé používané datasety jasně pojmenované, stejně jako HMM profily či transkriptomy, seznamy organizmů, o kterých je v práci také řeč. Obdobně by měly být dostupné i vytvořené skripty či programy a výsledky s nimi získané (třeba výsledky vyvinutého algoritmu pro hodnocení kvality predikce struktury).
Domnívám se, že kapitola Metodika a v širším slova smyslu informace o použitých dat, metodách a postupech jsou nejslabší částí práce a nemyslím, si, že by bylo možné výsledky na základě v práci uvedených dat snadno reprodukovat.
V kapitole Výsledky se autor vydává na cestu za nalezením proteinu Sam50, který se v proteomu *Giardia intestinalis* nejspíše nevyskytuje, což je úkol nezáviděnihodný. S velkým obdivem jsem sledoval jak autor zkouší i cesty velmi krkolomné s malou nadějí na úspěch. I díky tomu je však jeho závěr, že se Sam50 v genomu *G. intestinalis* nevyskytuje poměrně

přesvědčivý. V diskusi nenalezení homologa Sam50 pak autor zvažuje variantu, že existující metody nejsou dost citlivé, aby identifikovaly divergentní sekvenci. Nediskutuje však příliš další varianty – např. kvalitu genomu – v letošním roce vyšla nová genomová sekvence za použití PacBio technologie (https://www.nature.com/articles/s41597-020-0377-y), která může změnit predikované proteiny. Genom *Giardie* má také poměrně vysoký obsah GC párů, což může komplikovat detekci homologů. Nevím také, zda je prokázáno, že Tom40 u *Giardie* je umístěn v membráně a plní stejnou funkci jako Tom40 např. u kvasinky.
Vysoce oceňuji pečlivost autora s kterou ověřoval kvalitu nově vyvinuté metriky na hodnocení kvality modelů na kvasinkových proteinech, ale nepochopil jsem jak byla uplatňována kontrola distribuce dihedrálních úhlů.
Celkově práce obsahuje až neuvěřitelné množství výsledků získaných obdivuhodně širokým spektrem bioinformatických metod – u některých z nich si však vzhledem k detailu popisu nejsem jistý, zda byly udělány dobře. Moc se mi také líbila autorova schopnost propojovat bioinformatickou práci s prací experimentální v laboratoři – domnívám se, že tato kombinace bude do budoucna velmi cenná.
Text je po formální stránce velmi kvalitní s minimem nedokonalostí (např. odkaz na obrázek 20 se objevuje už před odkazem na obr. 19, což je matoucí), práce jsou správně citovány, formát citací je jednotný. Odkazy na obrázky v textu však pro mě byly matoucí ohledně obsahu, který obrázky následně ukazovaly.
Diskuse je vzhledem k množství získaných výsledků i k množství použitých přístupů relativně skromná a jen velmi málo jsou výsledky konfrontovány s existující literaturou a diskutována nejsou ani experimentální data o lokalizaci proteinů.
Celkově však Vít Dohnálek dokázal svou diplomovou prací, že si osvojil velké velké množství metod, zvládne zpracovat obrovské datové soubory jako Uniprot, umí vytvořit funkční protokoly z existujících metod i vyvinout metody nové a umí o tom všem sepsat srozumitelnou zprávu a proto jeho práci jednoznačně doporučuji k obhajobě. Navrhoval bych vzhledem k výše uvedeným nedostatkům hodnocení velmi dobře.

K práci mám následující dotazy:

V textu píšete, že β-barelové membránové proteiny mají vyšší mutační rychlost než cytosolické proteiny – je tomu tak i v transmembránových částech proteinu nebo jen ve smyčkách?

Z čeho vycházeli Buczek et al. při popisu, že *D. discoideum* má homology Tom20 a Tom70?

Nebylo by lepší testovat kvalitu predikce β-barelů například i na kvasinkovém Tom40 než jen na proteinech z *Trichomonas*, které jsou hodně divergentní?

U hodnocení kvality modelů ukazujete na straně 30 Ramachandranovy diagramy pro kvasinkový Sam50 a Cyc2 a tvrdíte, že Sam50 má lepší Ramachandranův diagram, a proto mu věříte více, i když skóre má horší – je rozdíl 1.3% pozic v nepovolených oblastech u modelů skutečně významný? V textu také zmiňujete, že jste hodnotili kvalitu alignmentu – jak?

Hledali jste Sam50 u Spironucleus salmonicida, kde podle obrázku 11 máte k dispozici proteom?

Jak jste vybírali druhy pro fylogenetickou analýzu na straně 35?

Jak jste definovali vlastnosti aminokyselin v tabulce6? Tyrozín obvykle nebývá označován jako hydrofobní, i když velká část jeho postranního řetězce hydrofobní je.

Zkoušeli jste srovnat strukturně dva modely Tim proteinu navzájem a s templátem? Který model je podobnější templátu?

Myslíte si, že Tom40 plní u *G. intestinalis* stejnou roli jako u kvasinky?

V Praze 7.9.2020 Marian Novotný