

Oponentský posudek na disertační práci
Mgr. Tomáše Pilčíka
**Mapování genů pro kvantitativní znaky a stanovování jejich epistází
u rekombinantních kongenních kmenů myší.**

Práce byla napsána v pod vedením školitelky doc. Marie Lipoldové, CSc. z Ústavu molekulární genetiky AV ČR, v.v.i. Je založena na třech publikacích v odborných časopisech, což způsobuje také přirozené členění textu. V rámci standardního členění (Úvod, Cíl práce, Materiál a metody, ...) jsou kapitoly věnovány vždy třem tématům, souvisejícím spolu poněkud volně – odpovídá to zaměření publikací. Dvě témata se vztahují ke genetice myší, třetí se zabývá kvalitou statistiky v článcích v biomedicínských časopisech.

Z hlediska statistiky a jejích aplikací je nejzajímavější kapitola *3.2 Pravděpodobnostní poměry při tvorbě RCS* a kapitola *5.6 Statistická analýza dat*. V první z nich jsou aplikovány poznatky z teorie pravděpodobnosti, založené na takzvané kombinatorické pravděpodobnosti. To je v pravděpodobnostních modelech aplikovaných v genetice nejčastější postup. Zde je použit běžný předpoklad nezávislosti jednotlivých událostí, což by možná stálo za krátkou diskusí.

Ke kapitole *5.6 Statistická analýza dat*:

Většina zde používaných procedur vyžaduje nezávislost pozorovaných jedinců. Podle mého, tak jak rozumím popisovanému sběru dat, je tento požadavek v takovýchto studiích těžko dosažitelný. Je pravda, že při rozsáhlých souborech dat slabá forma závislosti často příliš neovlivní asymptotické vlastnosti statistických procedur, přesto mi v práci chybí diskuse na toto téma.

Zde používaná verze NCSS (z roku 1992) je v dnešní době, kdy softwarové společnosti dodávají každoročně zdokonalované verze svých systémů, již poněkud zastaralá. Existuje sice novější verze NCSS, ale ani tu bych pro práci s takto rozsáhlými daty a složitými modely nedoporučoval. Moderní postupy v NCSS vesměs nejsou zahrnuty, tento software navíc neumožňuje tvorbu vlastních procedur. Proč autor nepoužil některý z aktivně se rozvíjejících novějších statistických balíčků? Mezi nejkvalitnější a nejoblíbenější programy dnes patří velmi výkonný a volně šířený program R, který umožňuje značnou volnost při tvorbě programů a nabízí širokou řadu knihoven volně dostupných na internetu.

V práci se pracuje s rozsáhlými modely obsahujícími mnoho (řádově sta) vysvětlujících proměnných. To je velmi netriviální problém, se kterým se v aplikacích potýkáme velmi často. Úkolem je navrhnout jednodušší model, který však bude obsahovat všechny podstatné vysvětlující proměnné. Autor používá postupu postupného vylučování proměnných z modelů. Zde je ale nutné bedlivě hlídat výslednou hladinu testů významnosti, protože vylučování je prováděno tolikrát, kolik markerů sledujeme (provádí se tedy takzvané mnohonásobné testování), a v kombinaci s vícekrokovou analýzou pak hrozí, že skutečná hladina spolehlivosti je mnohem vyšší než deklarovaná.

Alternativou by mohlo být budování modelu odspodu rovnou ve zobecněném lineárním modelu. Tím by na začátku nebylo nutné celkový model rozdělovat na podmodely a provádět víceokrové postupy.

V současné době již existuje bohatě rozvinutá oblast matematické statistiky zabývající se výběrem modelu v případě, že možných vysvětlujících proměnných je velký počet, ale dá se očekávat, že jen malá část z nich je významná. Tyto postupy se dají najít pod klíčovými slovy „sparse estimator“, „lasso“, „penalized regression“. V práci použité metody jsou přeci jenom staršího data a nereflektují moderní vývoj matematické statistiky a jejích aplikací.

Autor upozorňuje na problémy a nedostatky v používání statistiky v biomedicínských časopisech (a nejen v biomedicínských!). Jde o věc bezesporu potřebnou a užitečnou. Nedostatky tohoto typu většinou souvisejí se závažnějšími problémy v experimentech; špatný návrh experimentu, nevhodná volba metod, nevhodná interpretace výsledků. Zde nezbývá než doufat ve vývoj „poučení“ autorů a redaktorů některých časopisů.

Dorbné překlepy a přehlédnutí

Strana 6, 2. a 3. řádek zdola: Namísto desetinných čárek, jak je v textu běžné, jsou uvedeny desetinné tečky.

Strana 15, 9. řádek zdola: namísto výrazu *množinu* patří zřejmě výraz *množina*.

Strana 31, tabulka 11: ve slově *nepřítadné* chybí písmenko.

Nakonec oceňuji autorovo rozhodnutí použít pro sazbu disertační práce systému \TeX . Zejména čitelnost vzorců se tím značně vylepšila.

Daniel Hlubinka
23. září 2008