Softwarový projekt: DNA Micro Assemblies

Motivace

Mnoho léků předepisovaných při závažných onemocněních vykazuje značně rozdílnou účinnost u různých pacientů. Cílem *personalizované medicíny* je zpřesnit diagnózu, zefektivnit léčbu a v ideálním případě zcela předejít rozvinutí choroby. Jedním z nejnovějších nástrojů, jehož skutečný potenciál se teprve začíná rýsovat, je znalost individuální DNA sekvence pacienta. Zatímco sekvencování prvního lidského genomu trvalo 13 let a stálo přes miliardu dolarů, dnes, pouhých šest let od dokončení, je možné sekvencovat 300-násobně větší množství DNA ve třech dnech za cenu tisíce dolarů. Ačkoliv problematika je v hrubých rysech již zvládnutá, prostoru pro vylepšování je stále dost.

Wellcome Trust Sanger Institute je jedno z předních světových pracovišť zaměřených na sekvencování DNA a genomiku, hlavním tématem výzkumu je studium vlivu genetické výbavy na zdraví a nemoci. Zde se mimo jiné sekvencovala největší část prvního lidského genomu, který zůstal nepatentovaný a veřejně přístupný zejména díky vlivu WTSI.
http://www.sanger.ac.uk

Ve spolupráci s WTSI nabízíme softwarový projekt zaměřený na hledání genetických variant, jednu ze základních úloh genomiky. Projekt umožní podílet se na skutečném vědeckém projektu.

Úvod do problematiky

Sekvencování druhé generace (*next generation sequencing*) je značně troufalý přístup: molekula DNA se fyzikálními metodami roztrhá na malé fragmenty, které se pak sekvencují všechny najednou v masivně paralelním procesu. Výsledkem je velké množství krátkých *čtení*, až 3×10^9 krátkých sekvencí, každá o délce 150-300 bází (písmen A,C,G,T). Protože každý z fragmentů pochází z náhodného místa genomu, problém lze přirovnat k obrovskému jednorozměrnému puzzle s překrývajícími se dílky. Úloha je značně komplikována repetitivními úseky, které tvoří 37% genomu, a náhodnými i systematickými chybami při čtení DNA sekvencí. Naším cílem je určit jednobázové substituce a krátké inzerce či delece oproti referenčnímu genomu. První generace algoritmů pro určování těchto variant je spolehlivá jen pokud se nacházejí ve snadných oblastech genomu a zkoumaný vzorek se příliš neliší od referenční sekvence, ve složitějších úsecích je však procento chybně rozpoznaných variant řádově větší. Současné algoritmy zcela spoléhají na správnost namapovaných čtení, modelují pouze nezávislé náhodné chyby a zanedbávají chyby systematické plynoucí z nesprávně umístěných či chybně zarovnaných čtení. V projektu se pokusíme zlepšit úspěšnost modelováním problematických oblastí v souladu s očekávaným počtem haplotypů pomocí De Bruijnových grafů.

Popis projektu

Cílem projektu je vytvoření aplikace a podpůrných nástrojů pro rozpoznání a vizualizaci genetických variant. Vzhledem k velkým objemům dat (typicky až 300Gbp nebo 130GB komprimovaných dat na jeden zkoumaný vzorek) je důraz kladený na rychlost a paměťovou nenáročnost aplikace. Projekt zahrnuje následující úlohy:

a. Příprava reálných a simulovaných dat

b. Návrh a implementace algoritmu pro rozpoznání obtížných oblastí vyžadujících podrobnou analýzu

c. Konstrukce De Bruijnova grafu v těchto obtížných oblastech

d. Návrh a implementace algoritmu pro zpětné mapování DNA sekvencí na grafovou strukturu

e. Určení nejpravděpodobnějších haplotypů s přihlédnutím k očekávanému počtu chromozomů a odhad pravděpodobnosti chyby

f. Možnost vizualizace problémových oblastí s vyznačenými opravami

g. Vizualizace výsledků a srovnání úspěšnosti nové metody

Předpokládané složení týmu: celkem 4-5 studentů; 2-3 pro úlohy **a**, **f**, **g**, 2 pro úlohy **b**-**e**.
Projekt nepředpokládá žádné počáteční znalosti z oblasti biologie, znalost AJ je vítána.
Jazyk aplikace není pevně dán, předpokládá se C a Perl, OS Linux.

Kontakt

David Hoksza, KSI MFF UK (hoksza@ksi.mff.cuni.cz)
Petr Daněček, WTSI (petr.danecek@sanger.ac.uk)