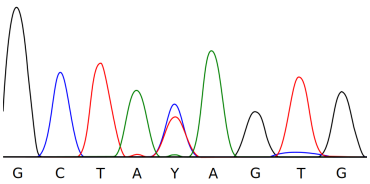
Přístup pro zpracování genomických dat sekvenovaných Sanger metodou

Kateřina Kratochvílová[1], Lucie Houdová[2], Pavel Jindra[3]

1 Úvod

Transplantace krvetvorných buněk je proces, při kterém jsou dárci odebrány krvetvorné buňky (štěp), které jsou následně vpraveny do těla pacienta trpícího hematologickou poruchou. K potlačení potransplantačních komplikací, jako je reakce štěpu proti hostiteli nebo návrat nemoci, je třeba vybírat co nejvhodnějšího dárce pro daného pacienta. Dárci se primárně vybírají podle shody v HLA genech. V současné době je prokazován vliv i dalších genů a dále se zkoumají vlivy jejich alelických variant.

Pro získání informace z DNA o konkrétních variantách sledovaného genu (alelách) se využívají sekvenační metody. Prezentovaná metodika vychází ze sekvenace Sanger metodou pro geny MICA/MICB. Výstupem této metody je chromatogram, což je posloupnost vln, kde každá barva značí jednu bázi. Cílem je určení posloupnosti písmen A, C, G a T (nukleové báze) označované jako sekvence. Pro zjednodušení, urychlení a zpřesnění analýz je snaha identifikaci automatizovat.



Obrázek 1: Chromatogram

2 Zpracování

Sekvence z jednoho jedince (pacienta/dárce) je metodicky získávána v pěti úsecích sekvence (tzv. pěti exonů). Každy exon je sekvenovaný obousměrně. Vzniknou tak tedy dvě sekvence - forward a reverse, které jsou vstupem do popisované pipeline (Obrázek 2). Sekvence je třeba sloučit a identifikovat na základě dostupných znalostí referenčních alel (v této práci využito IPD [*Robinson et al. (2013)*]). Výstupem jsou identifikované dvě alely, které mohou být obě stejné (homozygotní) nebo dvě různé (heterozygotní).

2.1 Pipeline

1. Získané sekvence všech jedinců pro daný exon jsou vloženy do jednoho *fasta* souboru spolu s referenčními sekvencemi pro daný exon.

2. Pro určení umístění exonu v sekvenci jsou sekvence ve vytvořených souborech zarovnány a seříznuty podle referenčních sekvencí. [*Sievers et al. (2011), Edgar RC (2021)*]

[1] studentka doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, e-mail: kkratoch@ntis.zcu.cz
[2] Západočeská univerzita, Fakulta aplikovaných věd, NTIS/KKY e-mail: houdina@ntis.zcu.cz
[3] Fakultní nemocnice Plzeň, Hematologicko-onkologické oddělení, e-mail: jindra@fnplzen.cz

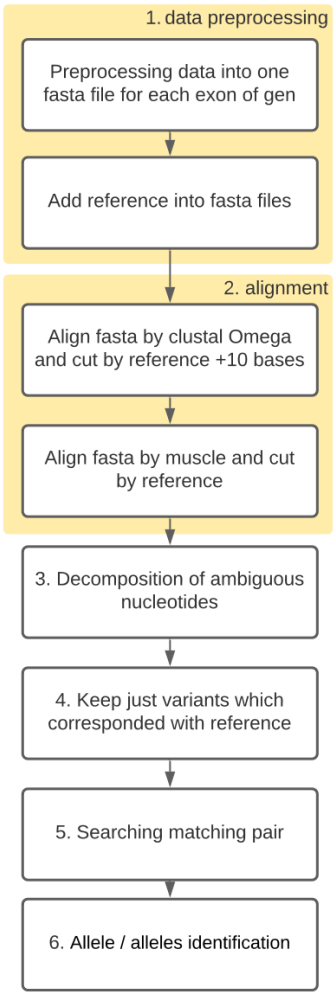
3. V sekvenci se mohou objevit nejednoznačné báze (Y, S, K a jiné). V chromatogramu je to možné pozorovat jako dvě vlny přes sebe. Toto písmeno v sobě typicky uchovává dvě báze. Rozdělení probíhá podle IUPAC klíče a to tak, že jsou sekvence, ve kterých se toto objevuje, rozděleny na dvě sekvence, kde v každé sekvenci je jedno z možných písmen (v případě Y je to C a T). [Cornish-Bowden A. (1984)]

4. Sekvence, které neodpovídají žádné referenci jsou odstraněny.

5. Pokud je sekvence určena jako heterozygotní, je zachován pouze shodující se pár (pár s odlišnými bázemi na neurčitých pozicích).

6. Alely jsou určené pomocí průniků výsledků na všech exonech.

Chybovost pipeline je způsobena podobností alel (rozdíl i pouze v jednom písmenu), chybou sekvenace, heterozygocií či neustále nově příchozími alelami. Pro odstranění části zjištěných nedostatků probíhalo předzpracování dvěma způsoby. Prvním bylo spojení forward a reverse sekvence do jedné sekvence pomocí nástroje CAP3 [*Huang, Madan (1999)*]. Nástroj se rovněž postaral o převedení chromatogramu do sekvence. Druhým způsobem bylo zpracovávání forward a reverse sekvence odděleně. V tomto případě probíhalo převedení chromatogramu do sekvencí pomocí Biopythonu.



Obrázek 2: Pipeline

3 Závěr

Navržená pipeline usnadňuje práci při zpracování genomických dat, avšak neřeší všechna úskalí. Stále je v některých případech potřeba manuálního zásahu. Do budoucna je snaha minimalizovat tyto zásahy za pomoci metod umělé inteligence.

Poděkování

Příspěvek byl podpořen grantovým projektem MZ ČR - AZV NV18-03-0277 .

Literatura

Robinson et al. (2013) *IPD The Immuno Polymorphism Database, Nucleic Acids Research.*

Sievers F et al. (2011) *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.*

Edgar, RC (2021), *MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping*

Huang X, Madan A. (1999), *CAP3: A DNA Sequence Assembly Program.*

Cornish-Bowden A. (1984), *A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res.*