Tutoriál 4: Eukaryotní geny

Pokud máme k dispozici nukleotidovou sekvenci určitého úseku DNA a chceme zjistit jestli kóduje nějaký gen, pak ji můžeme zadat do programu *BLAST* (konkrétně verze *blastx*) a prohledat známé proteinové sekvence. Nevýhody tohoto postupu jsou v podstatě dvě.

Zaprvé, program *BLAST* nám nalezne jako „hit“ nějaký protein, o kterém se můžeme domnívat, že má stejnou (nebo minimálně podobnou) funkci a podobnou strukturu. Co nám ale *BLAST* neřekne je informace kde sekvence kódující protein začíná a kde končí. Něco nám může napovědět translace nukleotidové sekvence do sekvence proteinové, ale úsek mezi dvěma stop-kodony nemusí odpovídat skutečné sekvenci proteinu. V eukaryotních organismech se navíc objevuje problém existence intronů.

Zadruhé, i když je to dnes nepravděpodobné, je možné, že kódovaný gen nemá v databázích vhodné homologní sekvence. Pak nemůžeme nacházet geny na základě podobnosti, ale můžeme je nacházet podle známých iniciačních sekvencí.

Nejprve si ukážeme postup jak procházet databázi sekvencí lidského genomu pokud známe lokalizaci genu. Databázi sekvencí lidského genomu můžeme nalézt na stránce http://www.ensembl.org (je možné se k němu doklikat i ze stránky http://www.ebi.ac.uk/). Databáze *ENSEMBL* obsahuje nejen lidský genom, ale i genomy dalších eukaryotních organismů. Pokud na stránce kliknete na obrázek sochy (býval to Michelangelův David, co je to teď nevím), tak se dostanete na lidský genom. Zde naleznete několik odkazů, které vedou například k hrubým sekvencím. My si na ukázku nalezneme gen lidské cystathionin-β-synthasy (CBS), o které víme, že se nachází na chromosomu 21 někde v oblasti 6,4 – 6,5 Mb. Proto klikněte na odkaz karyotyp. Tím se zobrazí schematický obrázek všech 22 somatických a obou pohlavních chromosomů. Chromosomy jsou vyobrazeny s pruhy tak, jak to odpovídá barvení při mikroskopickém vyšetření karyotypu. Gen pro CBS se nachází na chromosomu 21, tedy klikněte na něj a zvolte nejprve „Chromosome summary“. Objeví se vám stránka s vyobrazením chromosomu nastojato a spolu s četností genů, dalších nekódujících sekvencí (sekvencí předpokládaných, ale neověřených genů) a dalších elementů. Nahoře se objeví schema chromosomu 21 zobrazené na šířku. Začátek chromosomu zcela vlevo odpovídá počátku (0,00 Mb) a konec chromosomu zcela vlevo odpovídá konci chromosomu (cca 48,13 Mb). Abychom našli CBS tak musíme zoomovat v oblasti vlevo. Pak by mělo být možné se dozoomovat k odkazu „CBS“. Pokud na něj kliknete a kliknete na označení sekvence, tak se vám zobrazí seznam možných transkriptů. Klikněte na transkript `CBS-001`, čímž se objeví jeho schema. Pak můžete za úkol spočítat počet intronů, exonů, případně vzájemný poměr. Můžete kouknout i na odkazy *exon*, *transcript*, *protein* a *gene*.

Introny a exony CBS byly v případě databáze *ENSEMBL* identifikovány porovnáním její nukleotidové sekvence a sekvence cDNA nebo cDNA fragmentů (tzv. expressed sequence tags, *EST*). Nyní si ještě můžeme ukázat jak by introny a exony identifikoval automatický program. Můžeme si to vyzkoušet tak, že v záznamu genu CBS (nalezený v minulém odstavci) kliknete vlevo na Sequence. Sekvenci zkopírujte a vložte do programu *Genscan* (http://genes.mit.edu/GENSCAN.html). Program *Genscan* vám předpoví sestříženou sekvenci mRNA a proteinu. Pomocí programů, které znáte z minula, si můžete ověřit jak úspěšný byl program. Dále můžete zkusit eukaryotní verzi programu *GeneMark.HMM* (http://opal.biology.gatech.edu/) nebo program *FGENSH* (http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs& subgroup=gfind).

Identifikaci intronů a exonů podle sekvence cDNA si můžeme ukázat s použitím programu *Splign* (http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi). Tomuto programu můžeme zadat kód sekvence cDNA (jeho nalezení nechám na vás) a sekvenci genomové DNA. Místo genomové sekvence vyberte člověka a spusťte program.