Question 6

How would one retrieve the sequence of a gene, along with all annotated exons and introns, as well as a certain number of flanking bases for use in primer design?

doi:10.1038/ng971

This type of search can be initiated at the UCSC Genome Browser home page, located at http://genome.ucsc.edu. Select Human from the pull-down menu labeled Organism, and then click on Browser. This brings the user to the Human Genome Browser Gateway, from which a number of text- or position-based searches can be performed on current or older versions of the genome assembly. In this case, select the Dec. 2001 assembly, type the name of the gene of interest (PTPN1) into the position box, and then click Submit. The Browser returns all genes starting with the characters 'PTPN1' (Fig. 6.1). The gene of interest here is the one called PTPN1; click on the hyperlinked PTPN1 (arrow, Fig. 6.1) to view the genomic context of this gene (Fig. 6.2).

The text box at the top of Fig. 6.2 gives the absolute base pair position of this gene (chromosome 20, positions 48929540–49003636) and indicates that the gene spans 74 kb. The track labeled Chromosome Bands shows that PTPN1 is located at 20q13.13. Finally, the track marked Known Genes shows that the gene is on the forward strand, as the arrows on that track are pointing to the right. The exons within this gene are indicated by the vertical lines in the Known Genes track.

One way to obtain sequence upstream of a gene is described in Question 7. Here we explain how to retrieve flanking sequence on both sides of a gene. To retrieve an adequate amount of sequence with which to design primers, one can increase the size of the region displayed by changing the position numbers within the position box at the top of the figure. To add an additional 1,000 nt at the 5’ end and an additional 200 nt at the 3’ end, for example, change the text in the position box to 'chr20:4892854-49003836' and click Jump. This now redraws the graphic with the new boundaries.

To obtain the actual sequence within the region, click on the DNA link in the blue bar at the top of the page. This produces a new page, entitled Get DNA in Window (Fig. 6.3). Click the button next to extended case/color options and then click Submit. By selecting this option, the user can highlight features in the sequence by changing the format (case, underline, bold, italic) and/or color (red, green, blue) of the text. Colors can be varied in darkness and mixed together by changing the values in the boxes under Red, Green and Blue to any number between 0 and 255; examples of how to specify in RGB (red-green-blue) format color are given below the table. At this point, check the Toggle Case box in the Known Genes row, change the red saturation to 255 and leave the other color values set at zero (Fig. 6.4).

The Extended DNA Case/Color Options page can be used to combine and differentiate between genomic tracks. For example, return to the Options page, leave the Known Genes row as before but now also check the Underline square in the Mouse Blat row of the table. Clicking Submit produces a page on which the human exons still appear in red capital letters, but hits from the mouse sequence are now shown as underlined text (Fig. 6.6). In this section of the gene, the conserved mouse sequence overlaps with the exons.

One way to retrieve sequence for a defined chromosomal region at the NCBI is with the seq link on the MapViewer, visible when the Gene_Seq map is the master (Fig. 1.2). At Ensembl, export genomic nucleotide sequence with the Export→FASTA link in any ContigView window (Fig. 1.14, center yellow bar).
Figure 6.1

Known Genes

PTPN14 (also known as PTPN14, protein tyrosine phosphatase, non-receptor type 14), encodes a protein tyrosine phosphatase that is involved in signal transduction pathways. It is located on chromosome 17 and is conserved across multiple species.

mRNA Associated Search Results

HG0043 - Name: (clone labba-1L) non-receptor tyrosine phosphatase 3 (PTPN3) mRNA, complete cds.
HG0044 - Name: phosphotyrosyl-protein phosphatase (PTP1B) mRNA, complete cds.

Figure 6.2

UCSC Genome Browser on Dec. 2001 Freeze

Base Position

Track Controls

Note: Tracks with more than 100 items are always displayed in dense mode.
Figure 6.3

Get DNA in Window

Get DNA for chr20:48928540-48903835

[Checkboxes for options]

Figure 6.4

Extended DNA Case/Color

Extended DNA Case/Color Options

Use this page to highlight features in genomic DNA test. DNA covered by a particular track can be highlighted by case, underscore, bold, italic, or color. See below for details about color, and for examples.

Chromosomes (chr) Start (chr) Stop

Letters per line: [0-9] Default case: [U] Upper; [L] Lower

Track Name Toggle Underline Bold Italic Bold Green Blue

STJ Markers

Known Genes

Acacia Genes

Artemis Genes

Eqsensh++ Genes

Human mRNAs

Spliced ESTs

Unlabeled

Nonhuman mRNAs

Mouse Bait

Fish Bait

Overlap SHIPS

Random SHIPS

Represents

NCBI

Coloring Information and Examples

The color values range from 0 (darkest) to 255 (lightest) and are additive. The examples below show a few ways to highlight individual tracks, and their display. It’s good to keep it simple at first. It’s easy to make pretty but completely cryptic displays with this feature.

- To select some known Genes in upper case red bold, check the appropriate box in the Toggle Case column and set the color to pure red.
- RDB (255,0,0). Upon submitting, any known Gene within the designated chromosome interval will now appear in red capital letters.
- To see the overlap between known Genes and Genomic predictions by setting the known Genes to red (255,0,0) and Genomic to green (0,255,0) Places where the known Genes and Genomic overlap will be painted yellow (255,255,0).
- To get a level of coverage effect for tracks like Spliced ESTs with multiple overlapping items, initially select a darker color such as deep green.
- RDB (255,0,0). Nucleotides covered by a single EST will appear dark green, while regions covered with more ESTs get progressively lighter – or, you could use the image to indicate that the ESTs overlap.
- Another track can be used to mask unannotated features. Setting the Programaddress track to RDB (255,255,255) will white-out Genomic predictions of LACS but not the main Genomic features, masking with Known Genes will show what is new in the gene prediction sector.

Further Details and Ideas