EBI_01277 Technical Test

Please follow the instructions below and submit your response to hr@ebi.ac.uk. Please host your source code on a personal GitHub project space, or alternatively zip it and email the link to us. You may spend as much or as little time on this as you see fit but your code must compile and be able to be run in a single step. Please provide documentation on why you made the choices you did, how to run your application, and anything else we should be aware of.

Background biology

A **gene** is defined region of a genome that codes for one or more functional products. They are generally referred to by a symbol, for example BRCA2, and also a stable identifier, for example ENSG00000139618. A symbol is a short-hand easy to remember name for a gene whereas a stable identifier is akin to a primary key. A gene codes for one or more **transcripts** and each of these can code for a single **protein**.These transcripts and proteins also have stable identifiers and it is these that are the functional products of a gene. Chemically, both genes and transcripts are **nucleic acids** and their **sequences** are represented by a 4-letter alphabet (ACGT). Proteins are chemically very different from nucleic acids and their sequences are represented by a 20-letter **amino-acid** alphabet.

Ensembl resources

Ensembl has a well documented set of public RESTful APIs available from http://rest.ensembl.org that you can use to retrieve information about genes, proteins, their sequences and other major attributes.

One endpoint http://rest.ensembl.org/lookup/symbol/homo_sapiens/BRCA2.json?;expand=1 allows you to retrieve information about a gene by its symbol - in this example the species in question was *Homo sapiens* (human) and the gene symbol is BRCA2.

Another endpoint http://rest.ensembl.org/sequence/id/ENSG00000139618.json allows you to retrieve the DNA sequence of a gene as identified by its stable identifier. The endpoint can be given additional parameters such as *type* to control the type of sequence retrieved and *multiple_sequences* to allow the returning of multiple sequences from a single stable identifier.

1. Task

Your task is to build an application that takes in the following information:

- A gene symbol e.g. BRAF

- A position in a protein sequence e.g. 600
- An amino acid letter e.g. V

The application should display:

- All transcripts from that gene with the given amino acid at the specified position
- For each a link to the Ensembl website of the format http://www.ensembl.org/id/:id where :id is the stable of the transcript.

You should use the technology you feel most comfortable with but the solution should be a web interface.

2. Extending the Application

HGVS nomenclature is a way of representing variation within a sequence. An example is *ENSP00000419060.2:p.Val600Glu* which means there is a variation in protein ENSP00000419060 (version 2) at position 600 that changes a Valine to a Glutamic acid.

You should extend the interface from part 1 to allow a user to enter a HGVS string of the above format, parse it into its constituent parts, check the validity of the nomenclature (ie that ENSP00000419060.2 does indeed have a Valine at position 600) and have the interface return the same results as above. The application you submit for this extension must not have a ‘Submit’ or ‘Go’ button.

3. Testing

What strategies would you employ to test the application you have just written? Please consider how you would do this during both the development stage and also when application is live in production.