Spec for additions to Feature query - Subprotein level features

1 Definition

This will add to feature query, allowing for querying protein features.

This query will use name/aliases of protein domains that will be mapped from protein positions to genomic positions based on Uniprot annotation file and genomic annotations used in feature query for viral proteins to allow querying for annotated functional or topological domains or motifs within proteins.

These features will be added on top of the ones already implemented in feature query. In order the make simpler the query for the now too long list of features, for old as well as for new ones, we should implement a different UI approach, where UI shows dropdown menus to specify (select) type of feature (gene, cds, stem loop, ..then the new ones aded here: topological domain, region, ..), and specific types thereof when available (e.g. cytoplasmic, within topological domain), and allow filter by cds or mature peptide aliases to narrow down the search (select all as default will search for instances of the selected feature type across all genome).

2 Prerequisites

For the new features, here is a table with protein ids (Uniprot IDs) of gene product (i.e, cds), protein name, short name, aliases, feature type, start and end position of feature, feature name (specific instance) when available, feature characteristics/note, e.g.

1. ”P0DTC2”; ”Spike glycoprotein”; ”S”; ”E2, Peplomer protein”, ”Motif”, 1269, 1273, ”KxHxx”

2. ”P0DTC2”; ”Spike glycoprotein”; ”S”; ”E2, Peplomer protein”, ”Region”, 437, 508, ”Receptor-binding motif” ”binding to human ACE2”

There are 19 feature types annotated in the table, column *feature_type*: ”Signal peptide” ”Chain” ”Topological domain” ”Transmembrane” ”Repeat” ”Domain” ”Zinc finger” ”Nucleotide binding” ”Active site” ”Metal binding” ”Site” ”Motif” ”Disulfide bond” ”Glycosylation” ”Region” ”Modified residue” ”Binding site” ”Non-terminal residue” ”Natural variant”
If I was to choose just some of them to start, I will choose ”Chain” , ”Region” , ”Domain”, ”Repeat”, ”Active site”, ”Topological domain”

Note: ”Chain” will be somewhat similar to mature peptide already present in feature query, although this is a different query: while mature protein contains signal peptide and stop codon, chain will not. Also, it includes Spike protein chains, that are not annotated in mature peptides.

3 Query specifications

1 A filter by Gene/CDSs products or mature proteins (select all option by default, checkboxes for the desired ones). Here, include all cds and mature peptides by their name/alias as options.

I Gene/CDS product selection can be filtered directly in table using columns *cds_name*. Mature proteins, which come from one or both of the two polyproteins in CDSs ORF1a/ ab will be mapped upon query to one or both of these CDS using newly added column *locus_mapping* in table annot_coord_table.csv (the one being used so far in feature query). The corresponding CDS(s) will be then used for the search in DB, since the coordinates in *nt_coord_start* and *nt_coord_end* are based on these and not the mature proteins. However, the search for mature peptides will work as a ”filter” to narrow down the hits to those found between the mature protein genomic coordinates, as in annot_coord_table.csv.

2 A dropdown menu to select first the type of feature to be queried: this menu contains all unique values of *feature_type*, e.g. "Modified residue".

3 A Dropdown to select a specific subtype within feature type or the desired instance of feature, e.g. "Lumenal", "Receptor-binding motif" : values from *feature_name* column (select all as default)

I After the feature type is selected, the list in dropdown would show only the values corresponding to the selected feature type (I would imagine this menu would appear only in the cases needed, listed below:)

This is the list of feature types for which we should have this "feature_name/note" dropdown menu: "Chain", "Domain" , "Region", "Motif", "Active site", "Repeat" , "Metal binding" (for these values are subtypes, in case you want to add subtypes and names as a different dropdown menus) "Glycosylation", "Topological domain" , "Transmembrane" , "Site", "Metal binding", "Repeat", "Modified residue", "Zinc finger", "Nucleotide binding" (for these ones, they are actual feature names, identification of just one feature in the genome and not a group).

I After user selections have being used to filter columns in table, the resulting hit(s) will be searched in DB using the columns *nt_coord_start* and *nt_coord_end* to perform a region query of the region, including both.

4 Response

Response will be same as that of a `feature query`