Подсказки к заданиям по банкам данных

Часть I (содержимое записей)

1. (b) Авторы записи — в поле RA
   (c и d) Класс данных и раздел EMBL указаны в первой строке документа. См. списки классов и разделов в описании банка EMBL: https://ftp.ebi.ac.uk/pub/databases/embl/doc/usrman.txt, глава 3.
   (f) Координаты генов, в том числе генов тРНК, ищите в поле FT.

2. Описание формата Fasta см. в презентации или в Википедии: https://en.wikipedia.org/wiki/FASTA_format . Описание (description) последовательности не обязательно, но желательно; название (identifier) последовательности в данном случае не нужно делить на поля, но оно должно быть достаточно информативным.
   Для редактирования файлов (если не владеете программой Far manager) пользуйтесь Notepad++ или WordPad; в последнем случае следите за тем, чтобы сохранить файл в текстовом формате. Расширение файла (часть имени файла после точки) должно быть “fasta” (например, файл можно называть leutrna.fasta, но нельзя leutrna.txt или leutrna.fasta.txt).

3. В поле FT смотрите информацию о закодированных белках и находите нужные данные.

4. (d) Ссылки на другие базы данных, в том числе EMBL — в поле DR
   (e) Координаты зрелых белков — поле FT, ключ CHAIN
   (f) Варианты последовательностей в разных изолятах — поле FT, ключ VARIANT
   (g) Сведения о публикациях — в полях, начинающиеся с R (RA etc). Затем см. PubMed.

Часть II (поиск)

1. (Вариант 1) Работайте на сервере `https://www.ebi.ac.uk/ena` . В меню Search выберите “Advanced search”. В Data type выберите “Nucleotide sequences” и нажмите Next. В “Taxonomy and related”, выберите NCBI Taxonomy и в окошко впишите либо “Viruses”, либо (лучше) латинское название какого-либо семейства вирусов; поставьте галочку против “Include subordinate taxa”. В “Database record” выберите Dataclass = STD. В “Titles, aliases and descriptions” выберите Description и напишите “genome”. Найдите кнопку Search и нажмите её. Дождитесь завершения поиска и выберите запись, описывающую геном приглянувшегося вам вируса (внимательно, не все записи описывают полные геномы или хотя бы сегменты геномов!).
   (Вариант 2) На сайте `https://www.ncbi.nlm.nih.gov/` найдите в правой части гиперссылку Nucleotide, пройдя по ней, выберите Advanced. В форме поиска выберите Organism и в окошке справа наберите название какого-либо семейства вирусов. Нажмите Search, дождитесь завершения поиска и слева щёлкните по слову RefSeq, чтобы ограничиться только референсными последовательностями. В списке найдите подходящий геном.

2. (После поиска в ENA) Пройдя по гиперссылке слева от описания, найдите справа раздел Download и сохраните два файла: полную запись в EMBL-формате и последовательность в fasta-формате. Следите за расширениями имён файлов, они должны быть не txt, а embl и fasta. Если файл сохранился с неверным расширением, нужно переименовать его.
   (После поиска в NCBI) Пройдите по выбранной гиперссылке. Справа вверху найдите “Send to”, отметьте “File” и выберите нужный формат. Переименуйте файлы, чтобы у них были правильные название и расширение.

3. Гены и белки описаны в поле FT полной записи EMBL (Features полной записи RefSeq)

4. Найдите в описании белка в поле FT записи EMBL номер доступа (AC) банка Uniprot: как правило он состоит из шести символов, первый из которых — заглавная буква.
   Если AC записи Uniprot, к примеру, P00174, то полная запись находится по адресу `https://www.uniprot.org/uniprot/P00174.txt` , а в fasta-формате – по адресу

`https://www.uniprot.org/uniprot/P00174.fasta`.
В записи RefSeq ссылки будут на RefSeq Protein вместо Uniprot, файлы скачиваются аналогично нуклеотидному RefSeq (через Send to).

5. Откройте сайт Uniprot https://www.uniprot.org/ , пройдите по ссылке Advanced. В верхнем левом окошке выберите “Taxonomy [OC]”, в верхнее правое впишите название семейства, в нижнем левом выберите “Protein name [DE]”, против него впишите слово (слова) из описания белка. Получив список находок, поставьте галочки против 10–15 из них и нажмите Download, далее читайте и поступайте по смыслу.