**Задания 1**

1. Изучите поля записи Uniprot на любом примере. Help’ы - <http://www.uniprot.org/manual/>. Примеры – INS\_HUMAN (человеческий инсулин), HBA\_HUMAN – цепь А гемоглобина – найдете в Uniprot; в Uniprot можно искать и по ключевым словам. Пример запроса:  
   name:”RNA polymerase” AND organism:”Human”.  
   Открыв страницу записи переведите её в текстовый формат (кнопка “text”) – чтобы понимать как все выглядит на самом деле.   
   Составьте краткое описание белка в выбранном примере – как вы поняли аннотацию.
2. Упражнения по SRS <http://www.ebi.ac.uk/>
   1. Сколько записей, относящихся к белкам человека, в БД Uniprot? В БД SwissProt?
   2. То же, но с указанием complete proteome в поле keywords
   3. Какой из ответов лучше отражает число белков, закодированных в геноме человека?
   4. Сколько новых записей о человеческих белках поступило в Uniprot в октябре 2012?
   5. Найдите записи Uniprot, в которых есть ссылки на публикации Неифаха.
3. Сравните последовательности нескольких гомологичных (ортологичных) генов и их продуктов (белков) из двух видов бактерий одного рода. Опишите различия.
   1. Выберите по штамму из двух разных видов бактерий, принадлежащих одному роду, с известными полными геномами.
      1. (<http://www.ebi.ac.uk/> => genomes => ENA Genomes Server => Bacteria, в последней колонке должно быть слово Proteom, что значит, что белки описаны)

Выбирайте известные или интересные (… => Proteom или Google) вам бактерии!

* 1. Сохраните каждый из протеомов (в данном контексте – наборы последовательностей белков) двух этих бактерий в файлах.
     1. Используйте SRS (srs.ebi.ac.uk), банк Uniprot.   
        Запрос: поле Species - название штамма; поле Keywords – complete proteome.
     2. Используйте кнопочку i рядом с окошком чтобы узнать какие слова проиндексированы (если не получается).
     3. На странице запросе сформатируйте выходную таблицу так, чтобы присутствовали поля: ID, primary AC, Description, Gene name, Locus Name, Species, Sequence length
     4. Сохраните результат (=>Save: File(text), без ограничений на число записей, Save with view: что-то вроде UNIPROT1 – название той таблицы, которую вы создали)
     5. Повторите сохранение, на этот раз сохраните последовательности в формате fasta.
  2. Сравните полученные таблицы и отметьте строки либо с одинаковыми названиями (Description) белков, либо с одинаковыми именами генов (Excel, команда ВПР – удобный инструмент для этого; или напишите скриптик)
  3. Скачайте аннотированные полные геномы выбранных штаммов (SRS, банк EMBL; слова complete genome здесь пишут обычно в поле Description). Если у штамма более одной хромосомы или есть плазмиды (маленькие ДНК), то скачивайте все. Другой вариант действий – скачать геномы со страницы со списком бактерий.
  4. Изучите аннотацию генов и их продуктов в записи EMBL.
  5. Для каждой выбранной пары белков (“одних и тех же” белков из разных штаммов)
     1. найдите координаты их генов в последовательностях геномов. Используйте Locus Name для поиска.
     2. Вырежьте участок последовательности ДНК, включающий ген и фланки по 100 п.н. с 5’ и 3’ концов; в случае гена на обратной цепи возьмите комплементарную последовательность!  
        Обе последовательности поместите в один файл в формате fasta (seqret пакета WEMBOSS; tfm seqret или seqret –help для подсказок)
     3. Откройте файл в Jalview, раскраска – по нуклеотидам. Выровняйте их (меню tools => alignment => muscle)
  6. Постройте выравнивание последовательностей тех же БЕЛКОВ.
  7. Опишите свои наблюдения за выравниваниями пары генов и пары белков.

**Задания 2**

1. Изучите насколько полно секвенирована данная вам хромосома человека.
   1. Скачайте хромосому в формате fasta с сайта NCBI => Resources => Data bases => genomes => human … и там файл с именем hs\_ref\_GRCh37.p10\_chr1.fa.gz (chr1 – 1я хромосома, GRCh37.p10 – идентификатор сборки генома)
   2. Найдите число и длины “контигов”; в данном контексте под контигом понимается максимальный участок из букв A, T, G, C (не считая white symbols и цифр); и длины участков – букв n или др., - разделяющих контиги.
2. Найдите GC-состав (процент пар G:C в ДНК) хромосомы. (команда geecee пакета EMBOSS)