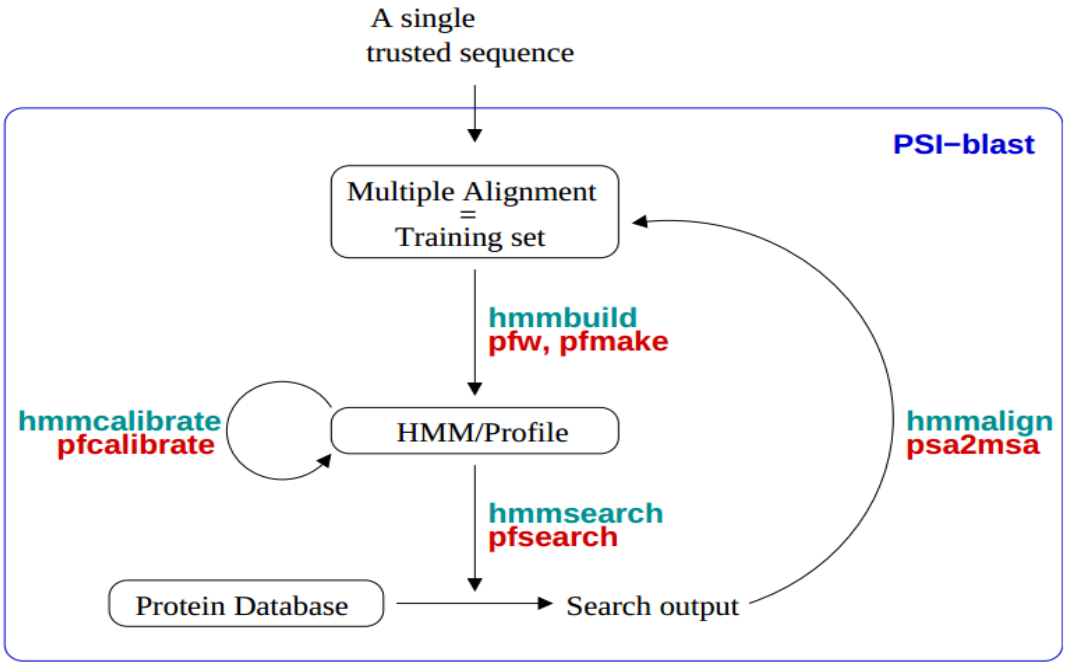
Lab 1

1. Скачать данные - Plasmodium falciparum chromosome 14
   или скачиваем chromosome 14 и ее анатацию с ncbi.
2. Прочесть позиции начал кодирующих регионов из gbk файла (CDS FEATURES). Позиции, в которых встречаются “<” или “>”, игнорировать(обозначают, что точные позиции начала или конца региона не известны).
3. Вывести в файл последовательности нуклеотидов для начал сайтов трансляции(+- 10 нуклеотидов от первой позиции трансляции, т.е. всего 21 нуклеотид). Неизвестные нуклеотиды ‘N’ не надо учитывать. Если кодирующий регион идет “не последовательно”, то надо считывать только кодирующую область. (Например, если кодирующая область join(1000..1008,1200..1500), то начальная позиция +10 будет соответствовать координате 1201).
4. Делим хромосому на 2 выборки (сайты трансляции, которые начинаются в первой половине хромосомы - обучающий набор, остальные - тестовый). Далее работаем с обучающей выборкой. Посчитайте частоту встречаемости нуклеотидов для каждой позиции сайта трансляции(PFM).
5. Используя найденные последовательности в пункте 2, постройте лого с помощью weblogo.
6. Посчитать матрицу весов(position specic scoring matrix), используя результаты, полученные в пункте 4.
7. Используя матрицу весов, постройте гистограмму scores для
   a. известных начал CDS в обучающем наборе (из пунктов 1-2)
   b. всех подпоследовательностей во всей хромосоме. Приблизить это распределение нормальным.
8. По данным, полученным в предыдущем пункте, постройте график FP, FN в зависимости от отсечки для score, по которой определяется, что последовательность в хромосомы - это CDS.
9. С использованием значением отсечки, которое кажется вам разумным, проведите поиск CDS во второй половине хромосомы и проанализируйте результат (FP, FN, sensitivity, specificity, and PPV).

Lab 2

PSI-BLAST, Generalized profiles, and HMMs



1. Цитохромы P450 – семейство наиболее мощных детоксицирующих ферментов организма. Последовательности для нескольких представителей представлены после условий заданий.
2. Выравниваем последовательности белков и берем в нем достаточно консервативный участок длиной 15-20 аминокислот. Для выравнивания можно использовать любую программу, например, http://tcoffee.vital-it.ch/apps/tcoffee/do:regular
3. По выравниванию(для выбранных консервативных участков) строим профиль, например, с помощью pfmake. (Объяснить, какую матрицу вы использовали и почему)
4. По базе UniProtKB/Swiss-Prot ищем белки, которые подходят к нашему профилю. (Для поиска по базе используется pfsearch). Какой cut off вы выбрали?
5. Найдите все цитохромы P450 в этой базе и проанализируйте насколько хорош найденный профиль (вычислить sensitivity, specificityand и PPV). Проделайте еще одну итерацию(начиная с пункта 3), используя найденные последовательности в качестве входных данных. Проанализируйте новые результаты. Стали ли они лучше/хуже и как вы думаете, почему? Что нашлось кроме цитохромов (посмотрите несколько последовательностей)?
6. Используя PRATT, найдите для изначальных цитохромов паттерн и прогоните его по базе UniProtKB/Swiss-Prot с помощью ScanProsite. Проанализируйте результаты.
7. *(необязательное) Прогоните psi-blast с одной из входных последовательностей. Какие последовательности нашел psi-blast?

>1cpt00
MDARATIPEHIARTVILPQGYADDEVIYPAFKWLRDEQPLAMAHIEGYDPMWIATKHADVMQIG
KQPGLFSNAEGSEILYDQNNEAFMRSISGGCPHVIDSLTSMDPPTHTAYRGLTLNWFQPASIRK
LEENIRRIAQASVQRLLDFDGECDFMTDCALYYPLHVVMTALGVPEDDEPLMLKLTQDFFGVEA
ARRFHETIATFYDYFNGFTVDRRSCPKDDVMSLLANSKLDGNYIDDKYINAYYVAIATAGHDTT
SSSSGGAIIGLSRNPEQLALAKSDPALIPRLVDEAVRWTAPVKSFMRTALADTEVRGQNIKRGD
RIMLSYPSANRDEEVFSNPDEFDITRFPNRHLGFGWGAHMCLGQHLAKLEMKIFFEELLPKLKS
VELSGPPRLVATNFVGGPKNVPIRFTKA
>1bvyA0
LNTDKPVQALMKIADELGEIFKFEAPGRVTRYLSSQRLIKEACDESRFDKNLSQALKFVRDFAG
DGLFTSWTHEKNWKKAHNILLPSFSQQAMKGYHAMMVDIAVQLVQKWERLNADEHIEVPEDMTR
LTLDTIGLCGFNYRFNSFYRDQPHPFITSMVRALDEAMNKLQRANPDDPAYDENKRQFQEDIKV
MNDLVDKIIADRKASGEQSDDLLTHMLNGKDPETGEPLDDENIRYQIITFLIAGHETTSGLLSF
ALYFLVKNPHVLQKAAEEAARVLVDPVPSYKQVKQLKYVGMVLNEALRLWPTAPAFSLYAKEDT
VLGGEYPLEKGDELMVLIPQLHRDKTIWGDDVEEFRPERFENPSAIPQHAFKPFGNGQRACIGQ
QFALHEATLVLGMMLKHFDFEDHTNYELDIKETLTLKPEGFVVKAKSKKIPLGGI
>1eupA0
ATVPDLESDSFHVDWYSTYAELRETAPVTPVRFLGQDAWLVTGYDEAKAALSDLRLSSDPKKKY
PGVEVEFPAYLGFPEDVRNYFATNMGTSDPPTHTRLRKLVSQEFTVRRVEAMRPRVEQITAELL
DEVGDSGVVDIVDRFAHPLPIKVICELLGVDEAARGAFGRWSSEILVMDPERAEQRGQAAREVV
NFILDLVERRRTEPGDDLLSALISVQDDDDGRLSADELTSIALVLLLAGFEASVSLIGIGTYLL
LTHPDQLALVRADPSALPNAVEEILRYIAPPETTTRFAAEEVEIGGVAIPQYSTVLVANGAANR
DPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLGIDADDVV
WRRSLLLRGIDHLPVRLDG
>1akd00
NLAPLPPHVPEHLVFDFDMYNPSNLSAGVQEAWAVLQESNVPDLVWTRCNGGHWIATRGQLIRE
AYEDYRHFSSECPFIPREAGEAYDFIPTSMDPPEQRQFRALANQVVGMPVVDKLENRIQELACS
LIESLRPQGQCNFTEDYAEPFPIRIFMLLAGLPEEDIPHLKYLTDQMTRPDGSMTFAEAKEALY
DYLIPIIEQRRQKPGTDAISIVANGQVNGRPITSDEAKRMCGLLLVGGLDTVVNFLSFSMEFLA
KSPEHRQELIQRPERIPAACEELLRRFSLVADGRILTSDYEFHGVQLKKGDQILLPQMLSGLDE
RENACPMHVDFSRQKVSHTTFGHGSHLCLGQHLARREIIVTLKEWLTRIPDFSIAPGAQIQHKS
GIVSGVQALPLVWDPATTKAV
>1cmnA0
APSFPFSRASGPEPPAEFAKLRATNPVSQVKLFDGSLAWLVTKHKDVCFVATSEKLSKVRTRQG
FPELSASGKQAAKAKPTFVDMDPPEHMHQRSMVEPTFTPEAVKNLQPYIQRTVDDLLEQMKQKG
CANGPVDLVKEFALPVPSYIIYTLLGVPFNDLEYLTQQNAIRTNGSSTAREASAANQELLDYLA
ILVEQRLVEPKDDIISKLCTEQVKPGNIDKSDAVQIAFLLLVAGNATMVNMIALGVATLAQHPD
QLAQLKANPSLAPQFVEELCRYHTAVALAIKRTAKEDVMIGDKLVRANEGIIASNQSANRDEEV
FENPDEFNMNRKWPPQDPLGFGFGDHRCIAEHLAKAELTTVFSTLYQKFPDLKVAVPLGKINYT
PLNRDVGIVDLPVIF