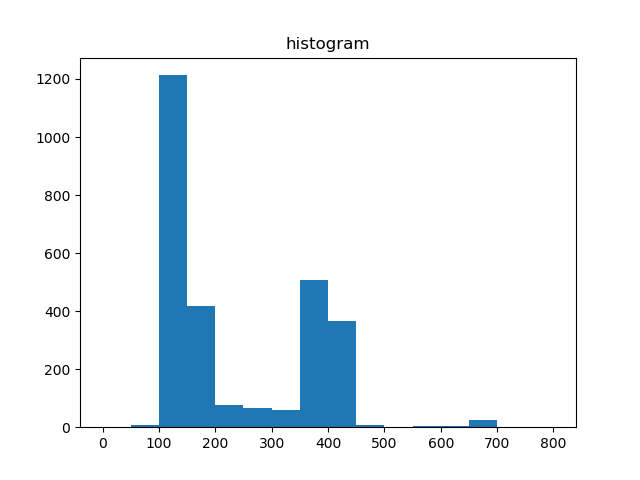
## 1.统计样本数

通过脚本对9大类的样本数据进行统计，得到样本总数2773个

## 2.获得正样本的长度分布

由于原来没有统计负样本，因此只有正样本，由BacHbpred的一篇论文，收集负样本尽量保持负样本的数量和长度分布尽可能一致，所以写脚本获得2773个数据的频率分布，获得大致的分布范围，便于查找适度长度的非氧结合蛋白



## 3.下载非氧结合蛋白

采用关键字uniprot:(proteins NOT name:potential NOT annotation:(type:"alternative products" isoform) NOT name:probable fragment:no NOT subunit NOT name:like NOT keyword:"Oxygen transport [KW-0561]" length:[x TO x] reviewed:yes) AND identity:0.9，获得去冗余后的数据，根据氧结合蛋白的分布，100-200收集1500个，200-350收集250个，350-450收集1000个.

## 4.生成pssm矩阵

原来的氧结合蛋白已经在服务器上跑完了，由于新增了2750个，所以需要添加新的fasta文件，重新在服务器上跑结果。