1. 根据交大的反馈，统计所有专利的FASTA文档中序列个数，与EMBL中的2700万条进行比对，看数量相差多少，若相差不大，说明数据没问题。

统计方式：统计下列18个文件包中fasta生物序列的个数。

“fastafiles\_emblrelease”文件下的文件：

em\_rel\_pat\_env.gz

em\_rel\_pat\_fun.gz

em\_rel\_pat\_hum.gz

em\_rel\_pat\_inv.gz

em\_rel\_pat\_mam.gz

em\_rel\_pat\_mus.gz

em\_rel\_pat\_phg.gz

em\_rel\_pat\_pln.gz

em\_rel\_pat\_pro.gz

em\_rel\_pat\_rod.gz

em\_rel\_pat\_syn.gz

em\_rel\_pat\_unc.gz

em\_rel\_pat\_vrl.gz

em\_rel\_pat\_vrt.gz

“蛋白生物专利序列FASTA”文件夹下：

epop.gz

jpop.gz

kpop.gz

uspop.gz