

## 应用密码学第一次作业解答

2. 求Hill密码的密钥空间大小。

**解：**问题本质是求整数剩余类环 $Z_m$ 上 $d \times d$ 的可逆矩阵的个数（特别地，这里 $d = 26$ ），将它们全体构成的集合记为 $GL(d, Z_m)$ ，容易发现 $GL(d, Z_m)$ 在通常矩阵乘法下构成群，这个群称为一般线性群。设 $m = \prod p_i^{n_i}$ ，根据中国剩余定理有同构

$$GL(d, Z_m) \cong \bigoplus GL(d, Z_{p_i^{n_i}}),$$

从而 $|GL(d, Z_m)| = \prod |GL(d, Z_{p_i^{n_i}})|$ 。为了确定 $|GL(d, Z_{p_i^{n_i}})|$ ，我们注意到对 $Z_{p_i^{n_i}}$ 上的矩阵，它可逆当且仅当矩阵的所有元素模 $p_i$ 后得到的矩阵是可逆的，这也就是说 $|GL(d, Z_{p_i^{n_i}})| = p_i^{(n_i-1)d^2} \times |GL(d, Z_{p_i})|$ （**代数学上的说法是： $Z_{p_i^{n_i}}$ 上的可逆矩阵是由 $Z_{p_i}$ 上的可逆矩阵“提升”得到的**）。

对于 $GL(d, Z_{p_i})$ ，注意到 $Z_{p_i}$ 为有限域，我们可以这样来确定其大小： $d \times d$ 的矩阵都是由 $d$ 个列向量构成的，这些列向量取自 $Z_{p_i}$ 上的 $d$ 维向量空间，我们要确定从这个向量空间有多少种方式可以取 $d$ 个向量能够线性无关。可以这样操作：先任意取一个非零向量作为矩阵的第一列，取法一共有 $p_i^d - 1$ 那么多（可逆矩阵不能有全零列）；接着我们取第二列，要求这一列一定与刚取的第一列线性无关，这样共有 $p_i^d - p_i$ 种取法（即从线性相关的向量之外再取）；接着我们取第三列，要求这一列一定与刚取的前两列线性无关，这样共有 $p_i^d - p_i^2$ 种取法（即从线性相关的向量之外再取，因为前两列生成2维子空间，共有 $p_i^2$ 个元素）；... 最后一列有 $p_i^d - p_i^{d-1}$ 种取法。因此有

$$|GL(d, Z_{p_i})| = \prod_{k=0}^{d-1} (p_i^d - p_i^k).$$

最终可得公式

$$|GL(d, Z_m)| = \prod_i p_i^{(n_i-1)d^2} \prod_{k=0}^{d-1} (p_i^d - p_i^k).$$

本题可以参考文献《On the Keyspace of the Hill Cipher》。有限域上一般线性群的大小公式是数学上一个基本的结论，对这个公式变形会发现它可以由 $q$ -阶乘（这是通常的阶乘函数 $n!$ 的一种推广）表出，这说明这个计数公式有重要的组合含义。



有兴趣的同学可以思考如下问题：

能否给出有限域 $\mathbb{Z}_p$ 上秩为 $k$  ( $1 \leq k \leq d$ )的 $d \times d$ 的矩阵的个数（可逆矩阵即为秩为 $d$ 的矩阵）？

3. 两段密文，第一段使用Substitution Cipher加密，第二段使用Vigenere Cipher加密，试确定它们的明文。

(1) 频率统计分析，统计单字、双字、三字母频率，结合英语语言特点猜测，连猜加蒙，没有固定的做法。最终可得明文：

I may not be able to grow flowers. But my garden produces just as many dead leaves, old overshoes, pieces of rope and bushels of dead grass as anybody's, and today I bought a wheelbarrow to help in clearing it up. I have always loved and respected the wheelbarrow. It is the one wheele...

这个题课件上给的密文其实少了一小段，给猜测可能带来不小困难，因为最后一句明文并未构成完整的英文句子... ..  

(2) 先使用Kasiski测试猜测密钥长度：观察密文中HJV字母段出现的位置，根据其间距的最大公因子，猜测密钥长度为6，进而可以通过计算该密钥长度下密文各列的重合指数，检验所猜测的密钥长度是否比较准确。确定密钥长度后为了确定密钥字，计算密文各列与自然语言的互重合指数确定该列的位移数；或者计算密文各列的互重合指数得到各列间的相对位移数，进而遍历某一系列的可能密钥导出其它列的密钥。最终可得密钥字为CRYPTO，明文为

I learned how to calculate the amount of paper needed for a room when I was at school. You multiply the square footage of the walls

by the cubic contents of the floor and ceiling combined, and double it. You then allow half the total for openings such as windows and doors. Then you allow the other half for matching the pattern. Then you double the whole thing again to give a margin of error, and then you order the paper.

Vignere密码的分析方法是非常深刻的、针对古典密码的分析方法，一些现代密码算法的分析也有类似的思想。它的分析思想可以这样理解：首先，单表替换密码（Caesar密码）之所以可以通过频率分析来破译是因为它的加密是由自然语言（比如英语）按简单规则替换的（移固定位数），所以从密文中随机取一些出来构成一个串 $S$ ，假设长度为 $n$ ，则 $S$ 与一个完全均匀随机的长为 $n$ 的字母串肯定有区别，原因就是自然语言中的字母存在频率的高低起伏，所以 $S$ 存在一定的非随机性。对多表替换密码（Vignere密码）而言，我们假设知道密钥长度 $l$ ，则可以把密文划分成 $l$ 个列，每一列都可以看成是由Caesar密码加密得到的。这种加密方式的特点是：明文同一个字母可能被加密成不同的密文字母，密文同一个字母也可能由明文不同字母加密得到，这样本来高低起伏的那个字母频率分布柱状图加密后频率的高低起伏被“削平”了，密文长得好像“比较”随机了，单纯想靠频率分布破译密文就不太可行，所以在大约100年左右的时间里Vignere密码都被认为是牢不可破的。但是，这种“比较”随机也都是来自对一个不随机的东西加了一些掩饰而已，仍存在一定的非随机性，如果我们能知道密钥长度，这种掩饰就被剥离了。如何获得密钥长度呢，有两种方式。一种是Kasiski测试，即观察密文中的重合字符段，看它们的间距，因为重合的密文段很可能是由相同的明文加密得到，他们加密完之所以相同是因为用了相同的密钥，所以这一间距体现了密钥的重复使用性，将这种重复使用性综合起来（计算最大公因子）就能估算密钥长度。当然这种方法比较粗糙，当密文比较短时可能行不通，但是确实能反映密钥长度的一些信息。

另一种方式是利用密文的统计特征发掘密文的独特性质，也就是说我们需要定义一个统计量，针对这个统计量而言自然语言、完全随机字母串、Vignere密码的密文序列三者的数值应该存在一些差异才行，并且这个

统计量应该能表面上体现、或本质上体现密钥长度这个我们想知道的值。什么是一个比较靠谱的统计量呢？只要回忆我们那个高低起伏的柱状图是如何被削平的就明白了，原来用字母的重合情况作为统计量则能够体现上述三个值的差异：用Vignere加密，相同字母可能被加密成相同，也可能加密成不同，不同字母也可能被加密成相同，这一切破坏了自然语言的固有的字母重合特征。因此我们可以定义重合指数，即一段字母序列中存在相同字母的概率。这个概率属于古典概型，简单的排列组合知识就能计算。假设对自然语言，这个统计量的值为 $IC_p$ ，对完全随机串值为 $IC_r$ ，对密文序列值为 $IC_c$ ，则 $IC_p - IC_r$ 体现了自然语言的非随机程度， $IC_c - IC_r$ 体现了密文的非随机程度而且这个值肯定比 $IC_p - IC_r$ 小(这是因为，如前所述，密文已经显得比较随机了)。这两个值的差异是什么造成的呢？是密钥长度！因为在不同的密钥字母下，密文不同列间字母的重合是导致密文变得比较随机的原因(不同明文字母可能被加密成了相同密文字母)。所以不严格地说它俩的比值应该就差不多是密钥长度了。当然这样估算比较粗糙，我们也有另一种稍严格的方法：对密文而言，其字母存在相同的概率（即重合指数）来自两个方面，一是从密文同一列来取相同字母，由于同一列是由Caesar加密来的，取得相同的概率乘上 $IC_p$ 应该就是这部分概率；另一方面是从密文的不同列取相同字母，这些相同字母可能是由不同明文字母加密得到的，所以取得相同的概率乘 $IC_r$ 应该就是这部分概率（乘 $IC_r$ 是因为考虑不同列时密文已经表现得随机了）。利用这个分解，我们也可以估算出密钥长度 $l$ （这是因为我们计算取出相同字母的概率时表达式中肯定会包含 $l$ ）。当然了，如果觉得这种估算也不靠谱我们还有终极方案：直接尝试可能的密钥长度（可以使用Kasiski测试先确定个大致范围）并遍历，对每次的密文分解计算各列的重合指数，发现各列的值都与 $IC_p$ 差不多时那个密钥长度应该就是我们想找的。有了密钥长度再想恢复密钥就比较容易了：一种想法是直接对各列进行频率分析，因为各列都是Caesar加密来的。但这种方式不一定能成功，因为我们按字母频率将某列密文字母恢复成明文字母，即便这种恢复是正确的，但由于这一列并不是一段自然的语言（想想我们这个密文矩阵是怎么排列的：应该逐行来读，行内是一个个的单词，

列内并不是), 如果我们不能正确恢复足够多的列, 或不具有超凡的语言水平, 或不是人群中万里挑一的人才, 想要成功恢复明文也不容易。靠谱点的办法是对每一列遍历其可能的移位情况, 检验它什么时候就是自然语言。这个检验用到的统计量是互重合指数, 它反映了密文列与自然语言的相关程度, 只有值最大时才能完全一致; 或者我们可以考察各个密文列的互重合情况, 从而找到它们密钥的间距, 进而通过确定一系列的密钥推导出其它列的密钥 (这是课上和讲义中介绍的方法)。

关于Vignere密码分析方法的`理解`, 可以查阅如下参考书P.11–P.16:  
Mark Stamp, Richard M. Low-Applied cryptanalysis: breaking ciphers in the real world, Wiley-IEEE Press (2007)