

Breaking a Vigenère Cipher

We are presented with the following substitution cipher:

ANYVG YSTYN RPLWH RDTKX RNYPV QTGHP HZKFE YUMUS AYVVK ZYEZM EZUDL JKTUL JLKQB JUQVU ECKBN
 RCTHP KESXM AZOEN SXGOL PGNLE EBMET GCSSV MRSEZ MXHLP KJEJH TUPZU EDWKN NNRWA GEEYS LKZUD
 LJKFI XHTKP IAZMX FACWC TQIDU WBRRL TTKVN AJWVB REAWT NSEZM OECSS VMRSJ JMLEE BMMTG AYVIY
 GHPEM YFARW AOAEI UPIUA YYMGE EMJQK SFCGU GYBPJ BPZYP JASNN FSTUS STYVG YS

Our first goal is to determine if it is monoalphabetic or polyalphabetic. We first do a frequency count:

a	14	g	12	l	13	q	5	v	10
b	8	h	8	m	16	r	11	w	9
c	7	i	5	n	13	s	18	x	7
d	5	j	11	o	4	t	15	y	16
e	22	k	14	p	13	u	14	z	11
f	6								

Notice all the letters appear several times, and the frequency does not vary much. This suggests a polyalphabetic substitution cipher. The IC is 0.041 for this cipher, so this suggests one also, and indicates the period is rather long (5 characters or more.)

First we seek to establish the period. We do a Kasiski examination, and write down all repetitions and how far apart they occur:

repetitions	first	next	interval	factors
YVGYS	3	283	280	2, 2, 2, 5, 7
STY	7	281	274	2, 137
GHP	28	226	198	2, 3, 3, 11
ZUDLJK	52	148	96	2, 2, 2, 2, 2, 3
LEEBMMTG	99	213	114	2, 3, 19
SEZM	113	197	84	2, 2, 3, 7
ZMX	115	163	48	2, 2, 2, 2, 3
GEE	141	249	108	2, 2, 3, 3, 3

The common factor to these is 2. But 2 occurs whenever the period is even, and is probably too short, so let us look at other factors. Possibilities are 3 (7 out of 8 intervals), 6 (6 out of 8), 4 (5 out of 8), 5 (1 out of 8), 7, 8, 9, 14, 16, and 28 (2 out of 8), and all others in 1 out of 8. 3 is probably too short, and 4 and 12 make the repetition of LEEBMMTG accidental, which is very unlikely. So the period is probably 6.

Working from this, we do a frequency count for each of the 6 alphabets. The following table summarizes the counts:

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
#1	3	1	0	0	0	0	1	3	1	1	3	3	7	0	0	1	1	0	4	2	4	3	7	0	0	3
#2	2	2	1	0	0	0	3	0	1	3	3	3	6	2	2	3	2	0	1	4	1	3	1	2	2	1
#3	1	1	0	0	5	1	4	1	0	0	8	1	1	4	0	0	0	3	2	3	2	2	0	2	4	3
#4	3	1	1	0	6	4	2	2	0	1	0	0	5	2	1	2	5	5	1	2	1	0	0	1	3	
#5	3	2	3	1	8	1	1	2	3	0	0	5	1	1	0	3	0	2	3	1	3	0	0	0	4	0
#6	2	1	2	3	3	0	1	0	0	6	0	1	1	1	0	5	0	1	3	4	2	1	1	3	5	1

To check ourselves, we compute the IC for each of the 6 alphabets:

#1	0.065	#3	0.061	#5	0.060
#2	0.041	#4	0.055	#6	0.052

Alphabets #2 and #6 have ICs that indicate they are polyalphabetic, with periods of lengths around 10 and 2, respectively. All the other alphabets have ICs that indicate they are monoalphabetic. So the measures of alphabets #2 and #6 are probably statistical variance, and we will assume that we are on the right track.

In what follows, the lower-case letters are the plaintext and the upper case letters are the ciphertext.

Now notice the counts for each alphabet. Three look like those expected of English, only shifted. For example, in alphabet #1, notice the long gap between N and R, which is surrounded by many letters in the ranges J to M and S to W. The normal alphabet profile has a similar feature, the gap being from V to Z, and the surrounding letters being R to U and A to E. This indicates that the cipher is a shifted one, and that S may be a. A similar gap (from D to H) occurs in the frequency chart of the second alphabet, so following the same reasoning, I is probably a. Substituting the resulting characters, we obtain:

```
ifYVG YalYN RptoH RDTsp RNYPd iTGHP prKFE YceUS AYenK ZYEhe EZUDt bKTUL rdKQB JciVU ECstN
RCTph KESXu sZOEN apGOL PofLE EBueT GCSan MRSEh eXHLP sbEJH TchZU EDecN NNRes GEEa dKZUD
tbKFI XplKP IAheX FACeu TQIDc oBRRL blKVN AroVB REioT NSEhe OECSa nMRSL reLEE BueTG AYdaY
GHPme YFARe soAEL chiUA YgeGE EMriK SFCOM GYBPr tPZYP rsSNN FalUS STgnG YS
```

From this point on, we can simply look for English words and constructs. The he in group 10 of the first line suggests the E in alphabet #6 is really a t; trying that out, and assuming again a shifted alphabet, we get:

```
ifYVG nalyN RetoH RDisp RNYed iTGHe prKFE nceUS AnenK ZYthe EZUst bKTUa rdKQB yciVU ErstN
RCiph KESmu sZOEc apGOL eofLE EqueT GChan MRStH eXHL e sbEJH ichZU EsecN NNges GEEa dKZUS
tbKFI mplKP IpheX FAReu TQIsc oBRRA blKVN proVB RtioT NSthe OECha nMRsa reLEE queTG AndaY
GHeme YFAge soAEa chiUA ngeGE EbriK SFrom GYBer tPZye rsSNN ualUS SignG YS
```

In the last group of line 3, And suggests and; also, note that in group 8 of line 1, the three letters nce suggest that the preceding one is a or e. Given these, most likely alphabet #5 is unshifted, so:

```
ifYVg nalyN retoH Rdisp RNYed iTGHe prKFe nceUS anenK Zythe EZust bKTua rdKQb yciVU erstN
Rciph KESmu sZOec apGOL eofLE equeT Gchan MRsth eXhle sbEJh ichZU esecN NNges GEema dKZus
tbKfi mplKp ipheX fareu Tqisc oBRra blKvn proVb rtioT Nsthe OEcha nMRsa reLEe queTG andaY
Gheme Yfage soAea chiua ngege ebrik sfrom gyber tPzye rsSnn ualUS signG Ys
```

In line 1, group 6, we see he again. Guess that the preceding letter, G, represents t; if so, and if the alphabet is shifted, the N should be a. We confirm this by looking in groups 2 and 3 on line 1. Group 3 begins with re, which suggests are, and indeed group 2 ends in N. Substituting,

```
ifYig nalya retoH edisp Rayed iTthe prKse nceUf anenK mythe Emust bKgua rdKdb yciVh erstN
eciph Krsmu sZbec apGbl eofLr equeT tchan Mesth eXule sbEwh ichZh esecN anges Grema dKmus
tbKsi mplKc ipheX sareu Tdisc oBera blKin proVo rtioT asthe Orcha nMesa reLre queTt andaY
theme Ysage sOnea chIha ngeGr ebriK ffrom Glber tPmye rsSan ualUf signG ls
```

At this point the message can be read off:

```
ifsig nalsa retob edisp layed inthe prese nceof anene mythe ymust begua rdedb yciph ersth
eciph ersmu stbec apabl eoffr equen tchan nesth erule sbywh ichth esech anges arema demus
tbesi mplec ipher sareu ndisc overa blein propo rtion asthe ircha ngesa refre quent andas
theme ssage sinea chcha ngear ebrie ffrom alber tjmye rsman ualof signa ls
```

The keyword is SIGNAL. Here it is, formatted as normal written English:

If signals are to be displayed in the presence of an enemy, they must be guarded by ciphers. The ciphers must be capable of frequent changes. The rules by which these changes are made must be simple. The ciphers are undiscoverable in proportion as their changes are frequent and as the messages in each change are brief.

-- From Albert J. Meyers' Manual of Signals.