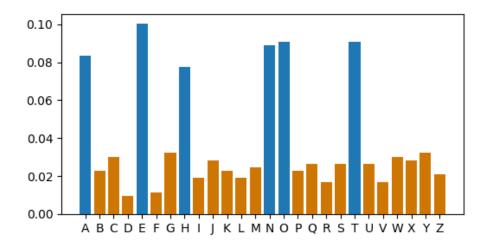
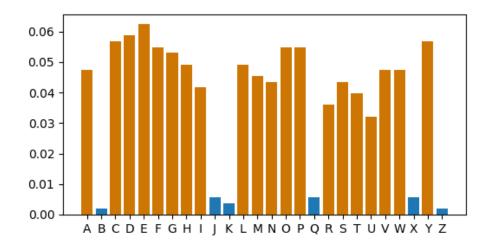
Unit 186 Attacking Purple with statistics and hill-climbing

In this unit we discuss how the use of statistics can help in analyzing a ciphertext that was encrypted with the Purple machine.

The weakest part of the Purple cipher machine is the division of the letters into sixes and twenties, and the handling of the sixes is the weaker of the two. Now, because of the split, the letters assigned to the sixes share their frequencies among themselves, while the twenties share their frequencies together. As an extreme example, suppose that we have chosen the six most frequent letters in English (A, E, H, N, O, T) as the sixes. When we encipher a typical piece of text with over 500 letters, we see the following distribution in the ciphertext. The sixes are clearly distinguished.



If instead we choose infrequent letters for the sixes, such as B, J, K, Q, X, Z, the distribution of letters in the ciphertext looks like this:

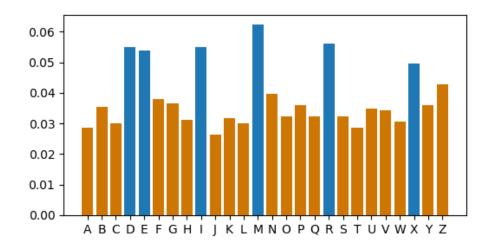


Here, the sixes are again easily distinguished by their low frequencies. Unfortunately, if the plugboard settings are chosen at random, the effect is washed out and it can be impossible to discriminate. But with a long enough text, it sometimes becomes possible again.

Here is a somewhat lengthy ciphertext with an almost randomly generated key (one—yes, just one—letter was changed in the selection of the sixes, and it makes enough of a difference). Let's see what we can do with it.

EFUZBKBWJVOVTVZWAQFRUFZBKXINIXIDOALXIZSUKRZHLMUFIIXWVNRVDGPNZEIK MOXMKYUBRMRJWYYEFXRMSGTSEBKSOXAXVKEYEERLOOBEJCUEIMXNPMWFLEJBMRRQ DPNSBZOXAGYZYODHPBTDVWCORNMTHXCZMS0IPTDEPRAWFMMRJKNSZIKSMYARLFPD DVRGCHNBODNUMUFPDDGCNTMMQWHGYQLUIFKRKTSYHMPMGPXQKDOGEYLXUUTGWPED WOWVBBPOMTKCTMMUOGDEBLVPREPMISKPGCRLCWTGFPIPWZYIXMYIDJEYHEGYCJLG EEGRYAQTQEDSCYJYNTHYEYOJZDDDYTSIUZWOQZENALNXCRHQPWKKHEIPPCNMRNEQ IIEYKAQDPAEVRUKGNNMJYRMVSMBFQQPRYIRGWZLJOXKNAZGOTSXDZSMMCRIMPUNX KZAAKCMMYVFPZUREMORGFODSJCBLZRYHUGEADMFMZUDATEHSOXZRXZDDYMAEJMMN TYICTIFDDAEHMHLEIOLDHXWKDZOUTPRBZVRRYFAKUCMODRDHFZLPMSTQMDNRREOJ INOURWYKVCEOEXQMJBENTTSSQJCIOMMEPNPIVCMYRMSMQRREFSIFJURVVGFGEPCR RNYZKLDAUMFILJTZDZDCQBYGTLVFXLNODYZYRHPADJZZUSFRKXZJBNCDFSYMZVJM WIFTDUMJDF0FHKLAVRBAPXIIFRR0BJQNQRTXWWS0IGNYEXXHGLNBNUZPYFVKJEUC EZO00ZDUERVEVTOSIBVMXTUGIXAOFNEOEJPPXRPKDNNEDCRSMMBHGWHEMFDZISKJ RTTXNIISXANQTLRTGKTWVULRNFVEKEIVSDDDUQMWXYILFLAKRIEXLVEINXPPCCOJ YQCFJFUORLHEGKZQQAVOIFKNTZDXRKEBMAAXYNSNFGVHFIHLQXEFGBCDEIOUXPNI EAIKEWBIYQMF0IOSRUPWLAIVASSUAMEXOMQHSASMPCNIXPZKVIXHRDXGGVIRWJGE RNFRPAVDCUHTEHIGVBCJJLRJZKXIWLRVOGNLIEBZMNGNKMMWLZHBDEILWSZDGXBC IGHSCMRCNFPYGFEXMEXREHLTGMGKUKEVYSDMMAILSOMBSPWFJRZVWBIZDDVESMST ETMDPLA0IILAYUBFANTRYZNXIIHLZXERNCVRBXVCIRHFAUMIJUVQGRIEFDTVSINQ ISUCOKDHYZWCDWHBEPXBURMHRFYMDYPBPMDDSMCVSINCOWUHXOFDMEYOCPJCVMOT GXXDHIZQBXZVHWGQRJBAYDZMENEBDYNRGXIDBVQCZXMVFJLBF0SXUZLXRMHUPDII YUWDGHDIMIRZLNDDFBHZMMDKXHZAXLTEJQGNWXUQQBKXUKSDLGDEEDXGOXRPFMWB LBZEIATERIBYKVUORUHEIRZHYKTGEIXBVRIEOPZYWFBWRJHBXOEVAOYFXYNOTMMA WMGMWWIZUZEPJVZZBQHDGGAMXZKXOPWFGGXZMXPXSPQLDIFIIBDURYFMOMUDXNCR LNWQADQGWNHDMGTIRFKRIXDPIJKZXWNFJYWCVEUEAGYOSDCQFTXROZMURZNMXDXB CLHBUDUGPKSWRBPEVNMRNIHMMDDSMHBJIBCX

Indeed, we can pick out the sixes:



The letters in the sixes have a wide range of frequencies in English, from seldom-used X to most-used E. This variation may help us in the next stage—finding the setting of the sixes switch.

The sixes switch advances by one for every letter that is enciphered, so let us bin the sixes letters in the ciphertext into twenty-five buckets, each for one position of the switch.

position in text	ciphertext
(modulo 25)	letters
0	DDDEEEEEEMRRXXXXXXX
1	DDDDDDDDEEEIIIIIIIIIIIRRRX
2	DEEEEEIIMMMMRRRRRRRRR
3	DDDIIIIIIIIIIMMMMX
4	EEEEEMMMMMMMRRRRRXXXX
5	DDDDDEIIMMMMMRRRRRRRRRRR
6	DDDDDDDEEIIIIMMXXXXXXXXXX
7	EEEEEEEEEIIIIRRRRXXXX
8	DDDDDDDDDDEEEMMMRXX
9	DDDDEIIIIIMMMXXX
10	DEEIIIIIMMMMRXXXXXXX
11	DEEEEEEEEEIIIRRXX
12	DDIIIIIIIIMMMRRRRRRRRRRX
13	DEEEEIIIIIMMMMXXXXX
14	EIIMMMMMMMMRRXXXXXXXXXXX
15	DDDEEEEEMMMRRRRRX
16	DDDDDDDDDDIIIMMRRRRRXX
17	DDDDEIIIIIIIIIIMMMMRR
18	EEEEIIMMMMMMRRXXXX
19	DDDDDEEEEIIMMMMRRRRRR
20	DDDDDEEIMMMMMMMXX
21	DDEEEEIRRRRRRRRXXX
22	DDDDDDDDDDEEMRRRXXXXX
23	EEEEEEEIIIMMMMMMRXXX
24	DDIIIIIMMMMMMMMMMRRRRRXX

Notice that some letters are missing, such as I in bin 0. It is likely that I is the encipherment of X by the sixes switch for that bin, and since X is very infrequent, I is missing in that bin. But we can be more general than that, and consider all the letters of the sixes. Convert the above lists of letters into frequencies in each bin:

position in text			frequ	encies		
(modulo 25)	D	Е	I	М	R	Х
0	0.15	0.3	0	0.05	0.1	0.4
1	0.28	0.12	0.44	0	0.12	0.04
2	0.0455	0.2727	0.0909	0.1818	0.4091	0
3	0.15	0	0.55	0.25	0	0.05

4	0	0.2174	0	0.3478	0.2609	0.1739
5	0.2	0.04	0.08	0.24	0.44	0
6	0.3077	0.0769	0.1538	0.0769	0	0.3846
7	0	0.4783	0.1739	0	0.1739	0.1739
8	0.5238	0.1429	0	0.1905	0.0476	0.0952
9	0.25	0.0625	0.3125	0.1875	0	0.1875
10	0.0476	0.0952	0.2381	0.2381	0.0476	0.3333
11	0.05	0.6	0.15	0	0.1	0.1
12	0.0833	0	0.3333	0.125	0.4167	0.0417
13	0.0526	0.2105	0.2632	0.2105	0	0.2632
14	0	0.0385	0.0769	0.3846	0.0769	0.4231
15	0.1765	0.2941	0	0.1765	0.2941	0.0588
16	0.4583	0	0.125	0.0833	0.25	0.0833
17	0.1818	0.0455	0.5	0.1818	0.0909	0
18	0	0.2105	0.1053	0.3684	0.1053	0.2105
19	0.2381	0.1905	0.0952	0.1905	0.2857	0
20	0.2778	0.1111	0.0556	0.4444	0	0.1111
21	0.1053	0.2105	0.0526	0	0.4737	0.1579
22	0.4783	0.0870	0	0.0435	0.1739	0.2174
23	0	0.3913	0.1304	0.3043	0.0435	0.1304
24	0.0769	0	0.1923	0.4231	0.2308	0.0769

The frequencies of the sixes letters in American English are

D	0.03961
Е	0.12497
I	0.07303
М	0.02541
R	0.06136
Χ	0.00198

The action of the cipher for each bin is a simple substitution on six letters. That substitution is determined by the portion of the plugboard that sets the sixes and by the position of the sixes switch. If the initial position of the sixes switch is s, and the permutation by the plugboard is π , then the true substitution for bin n is

$$S(n) = P^{-1}(\pi) S_6(s+n) P(\pi)$$
 (186.1)

where s+n is evaluated modulo 25. There are 6! = 720 possible permutations and 25 possible initial positions of the sixes switch, so there are 18,000 configurations that we have to check for the portion of the Purple machine that handles the sixes. Because we have computers, this is not a large number. We can score a pair (π, s) for bin n by shuffling the English frequencies in a way determined by Equation 186.1 and comparing the result to the frequencies observed in that bin. The comparison can be made by looking at the cosine of the angle between the sets of frequencies in a six-dimensional vector space (see Unit 9). For example, let's score the permutation $\pi = (2 \ 1 \ 3 \ 5 \ 6 \ 4)$ and switch setting 7. When a D enters the machine, it is mapped to the second line to the sixes switch. From Table 184.1, we see that

for setting 7 it emerges on line 4, which is mapped by π^{-1} to R. Likewise, E is enciphered to X, I to M, M to E, R to I, and X to D. The vector of frequencies is shuffled to

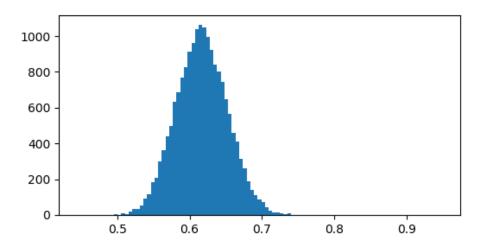
$$(0.00198, 0.02541, 0.06136, 0.07303, 0.03961, 0.12497)$$

In terms of permutations,

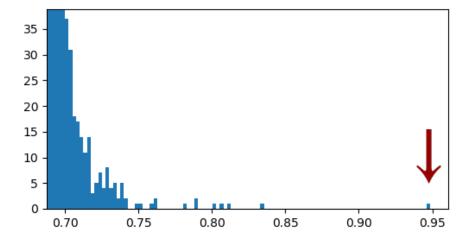
$$(2\ 1\ 3\ 5\ 6\ 4)^{-1}$$
 (5 4 6 3 2 1) (2 1 3 5 6 4) = (2 1 3 6 4 5) (5 4 6 3 2 1) (2 1 3 5 6 4)
= (6 4 5 3 1 2)

so the first number in the vector is was originally last in the list of English frequencies, the second was originally fourth, etc. Take the scalar product of that vector with this vector from bin 0 in the table:

Divide the result by the product of the lengths of the two vectors to get the score: 0.7478. The score for a pair (π, s) can be taken as the average of the scores of the twenty-five bins. A score close to 1 is considered a good match to the data extracted from the ciphertext, and may mean that we have found the true configuration of the machine so far as the sixes are concerned. We generated a score for every (π, s) , and here is an histogram of the results:



It is difficult to see the winning score, so here is an enlargement of the tail of the distribution:



The winner is π = MRIXDE and s = 3, with a score of 0.9484. The average is 0.6168, and the standard deviation is 0.0358, which puts our winner more than nine standard deviations above the mean. This gives us a high confidence in the correctness of the result.

At this point, if we want to, we can decrypt all the sixes in the ciphertext. We might even see the hints of words in there.

```
I.....R...E..IDE...D.I....ER.ERMI...ED....R...E..IDE...D.I....ER.E.
DI....RRI....I.EDR....R.M..E.IDE....E...ERDI..I...I..EDI.I...
...E.....R...E....E.E..R..I..I..ED.I...RED.....R..D..E...ER.I...
...E..E...ER....ER.....D..E.....R.I..E..E..D..I......ID.......
E....ME...MI....EM.RE.....I.....ERI.IE.D..R.....RI.M.....I
D.....E....E....IDE...R.....I....E.I......ED....EE.E.ERE....E..D.
E.....I..ED..D.....I.ED......E......IX..DE.REE...I......EI......
ERE.R...I.....E....E.RI.M..D..I...I.IE.DI.I.E.DI...D..E.RI.M.E..R
D.....E.IDE...D..E.RI.M.ERE..R...E....E..RI.....D..E.R....I.E.
.....R...R...E...I...D.....E.I....I...E...R.M..E.I.D......E...E
RM.DE.....E.I....E...ERE........E..I....M.DE.I....E..ME...
ER....E.I...RE..E..ED.R.MI.....EE.E.E...D..E.RI.M.....E....E..
I.D.R..E.....I...MI....ERE..E..ED.R.M..E..E..I..I...I.....E
ED.E....E...ER....EE.EMI...MI...EI..E...I....E.I....E.I....ER...D.
....RE..E....ME.....ERE....E.E..I....EI......RDEREDI....D....I...
ERE.R...I.....E....E.RI.M.E..R.ED....RD.....E...ERM...EEM...E
.I..ED....RD....ERE.R...I..I....E....I...E.I..ED.I..ER....ERE.R
...I.....I..RED.....I..ERE.R...I....E...E.RI.M.E..R.EDD....
.RD....E...ERM...EEM...E..RRIED...ER....ERE.R...I..I..I....E....
.I...E..RRIED..ME..I.....ER..ERE.....I..RED......ERE..REI.......
.ED.E.I..I.E.IR..M.....E.....ER..RE..ERRE.R...I......E.I....I...
..ME..R.M..ERED......D.....E..E..EI.M.RERE.R...I..E
```

Now things get harder. The sixes switch had only one choice for its configuration, and the portion of the plugboard relevant to the sixes was small. But the twenties switches have $3! \cdot 25^3 = 93,750$ possible configurations. And there are 20! (a big number) possible ways to configure that part of the plugboard. So we are going to resort to hill-climbing. The algorithm is almost identical to the one in Unit 36, and we will use tetragram fitness as the function to maximize. The differences are that we are only varying twenty letters of the key, that we have to try climbing as many as 93,750 hills, and that 1% of the time we are going to allow the routine to take the step even if the fitness is as much as 5% lower (we introduced this idea in Unit 90 when attacking a transposition cipher). The implementation is very fast, but checking all 93,750 configurations is slow, but we eventually find success when the configuration is

```
sixes = 3, left = 24, middle = 7, right = 12 fast = middle, medium = left, slow = right
```

The hill is climbed and at the peak is the twenties alphabet: AQJYLZCKSPOHGBVNUWTF. The full key for our example is

sixes = 3, left = 24, middle = 7, right = 12 fast = middle, medium = left, slow = right plugboard: MRIXDE AQJYLZCKSPOHGBVNUWTF

The plaintext is from *Opticks*, *or*, *a Treatise of the Reflections*, *Refractions*, *Inflections*, *and Colours of Light* by Isaac Newton:

I took a black oblong stiff Paper terminated by Parallel Sides, and with a Perpendicular right Line drawn cross from one Side to the other, distinguished it into two equal Parts. One of these parts I painted with a red colour and the other with a blue. The Paper was very black, and the Colours intense and thickly laid on, that the Phænomenon might be more conspicuous. This Paper I view'd through a Prism of solid Glass, whose two Sides through which the Light passed to the Eye were plane and well polished, and contained an Angle of about sixty degrees; which Angle I call the refracting Angle of the Prism. And whilst I view'd it, I held it and the Prism before a Window in such manner that the Sides of the Paper were parallel to the Prism, and both those Sides and the Prism were parallel to the Horizon, and the cross Line was also parallel to it: and that the Light which fell from the Window upon the Paper made an Angle with the Paper, equal to that Angle which was made with the same Paper by the Light reflected from it to the Eye. Beyond the Prism was the Wall of the Chamber under the Window covered over with black Cloth, and the Cloth was involved in Darkness that no Light might be reflected from thence, which in passing by the Edges of the Paper to the Eye, might mingle itself with the Light of the Paper, and obscure the Phænomenon thereof. These things being thus ordered, I found that if the refracting Angle of the Prism be turned upwards, so that the Paper may seem to be lifted upwards by the Refraction, its blue half will be lifted higher by the Refraction than its red half. But if the refracting Angle of the Prism be turned downward, so that the Paper may seem to be carried lower by the Refraction, its blue half will be carried something lower thereby than its red half. Wherefore in both Cases the Light which comes from the blue half of the Paper through the Prism to the Eye, does in like Circumstances suffer a greater Refraction than the Light which comes from the red half, and by consequence is more refrangible.

The last thing we should mention is that if we cannot pick out the sixes at the beginning of our analysis, then we need to hill-climb the entire plugboard. This does not add much to the effort.

Reading and references

Wes Freeman, Geoff Sullivan, and Frode Weierud, "Purple Revealed: Simulation and Computer-Aided Cryptanalysis of Angooki Taipu B," *Cryptologia* 27:1 (2003) 1-43, DOI: 10.1080/0161-110391891739

Programming tasks

- 1. Write a script or program that tests the 18,000 possible configurations for the sixes and finds that portion of the plugboard and the initial setting of the sixes switch. This assumes that we have been able to identify the sixes.
- 2. Write a script or program to perform the hill-climbing attack on the portion of the plugboard relevant to the twenties. Use tetragram fitness as the function to be maximized. Remember that this attack is performed for each configuration of the twenties switches (initial settings and speeds).
- 3. Write a script or program to perform the hill-climbing attack on the entire plugboard, for the cases in which we are unable to clearly identify the sixes.

Exercises

1. Give it a try.

WJQOITWTVPHJCNTYVTLZFHJQBTTQUWYFTHEQPXQIYWWCYWESSUEQSKLROUOCNONY XFEADSNFJSLHTPYTTGYNICGQCXJRTLIAXSANSKQNJRUGYHGPECHMDXDUTRSLECHA HNDRYYCHGJKYITMNQCAQRXCJWQLYXJAXCECSWNRCNOGVCSYDQLSKWXWAZAILWYMP PRJALPTYOLUORCDLDUOEXRRCCPFCARHQPOCQIPXDEPSGYJZDADNONSDDNCSPXHDW VWCCUIEJWWIODYJVTXCJASTIVPAPAHCRHSJJODJGRUSWXUSNJEHVNSCWTWLJTEUG NXXUWREQBYHCPNP0ICTLCAURQJXIATLSZUSAFJNWDBWL0EEPXYPHUYEJESDXPXWP LICTZRNCRONESVAPOHVXTIIXTOWLGYQLIUDZIKDJGCFYYCOYNADRXYELSSARWGNI NNBTDPNTIIETGCYHIHIRPPYEDAUATHOPMSPMIJRLIRTDCHEIVXWPHATTVYXNJHDO EQHIPCOHGDNSNJXDBUKLTPNMHYLWOSOUOCPRGWOQHIRHCLCUDDETTPWLVNPHPJSD HOJAGQRHRMOLCBPDONQAAHNUUTCWDWJMRCPWEJVJCJDRCHLGIQQKGRJDLLTDNVYT WXEAWFLDNIPDWYKWYSEPVHJSGSYTPRWPEROYRTWLXDRIOJSIJGSASRCXGEXYCWIA HJFYXGLLGYWOJQGOVQQAJAXSOIGNAGRZIAENFVWJNZTAENYPYCLCXXPIAPLQHMOB NRWPPKTXMAKGROETCGIIDCBTEJHEOGUFRHTGSMWYEOLPYTUMPEAEFEYRGWPTUYAL XBLOQLFARPHUYIGIHGTJAWGONJGLIJNATOMSQWSSRXOCWASEJRTXFNKFXIIXUDDN XITBDXPPOGYYHJXPHUXJWWAUUGTMCREOHEUZIINOJCDHPNUZAXTXXIUPGVXPNDAE DUUENGWJDGUGTAOCJHGMSECAANXTIECACCSDYPISASEDGILUOPEWGGCYJOFDHEJI FCDSIALCOQISSUCPIDXTHINLIXYULWZQQYRERITANWNGOWTWNTRSNSTCPQLIOVHH CZQUORJKXLDDSYNLINPYQJCDWDEHPTPAWSEANLAJYCKKOINBAUZEXHWOFCEYLLJX SQQDULBRSDPCNWLUTAOXRWIHHTDUSHLISWFXPPQHENDCPPGUTRLEVTLAZZHQONTA ASAGINYQYALHTPQIDLOSCOGPAVLIFWIQYWLDCSXRQXAXOGRSOJHGCINKEOSYYHUC TSRCNXERESNXTUAPALGPEECWJXLWKPICAMDQQNPCHTAILVNLVSAWGUCRPOLSYSZG LWUYUWLMIIHNNUSCJLXAUWJWPNALIMIIUAQRDFQOHYOSULECUPBNAGJAYNODSEXQ PPTRPMGJYZJESSORTOPOUOJSNOTCYCYGOINVIPAIPITPYPRGAZPANLAPOEPSCHDE UNFQRSXIACHPOTWLQPQHPWANIDSHAAXUWJUCAAUTPEDDBADNHPKHWIGIWENXDOSO IEJLIRIRIEAEEOOTRHIUFWHRICHESDSUTDHXVKOGDGCRNHUJWCVAOLOITCYGWJAP AICDVRYTOTURAPNGHAGDRWYJVPABDEJHAHYOCITNRGQGQIQRRLNTDVRUAYEWELKZ OJUXOQPPNXSXWGJFICMUXLYVDCPSXHTYOYGOHMPNXEDORYGGIX

2. Now another:

AAVHHBJWBYVYYJPWYYQSKAYALUDDTFBCHNLROWGHYYTTBMJIGHMXFHAOMRLWEYPP LEJWAGBKAPDPWRLYHERVHIIYNRZTYSBBHAWNPNITVAAAZYYLNIEIHFFIMHAMBLJQ VTNHVBXFSRVXWCJQHQUSUSTHORXRUPHTNUKCBOFCHRDCHLTGLTIQEOLQRQWENHYF JUINBAGBNWMFSFITHSOAFYJLBPRFNWATVQVKUCWBKCOQIIRREMDVWJMFVEREHMVU

DHGQCTTJLBUQNOZAHUCZSWLTFCRWAYKCQHZNHOMVTTZXTEFOTWNJAVAVTXAMCZAO DGJMHAUYWJWNYBZNYZKHGQUYTNZTSQFNVODWTJQSHNBYRLRKJUAYLVIRIVAALCIW HQOFNHGAPRDJRNMTTOHTWSTTGUWTTEVHIWHJRTTQXJWSWWRARIIPOOVZAWRHBWGH AMAHMXKAMGJNVNVNSAWHFUIHGSTANZWJLRDTSVWUNHWQNAMBAOGUPFBYMGNWPRHH BCAHKYDCQATAYXXPMCKRTVDJODTUMRTBHMNYNLWNXHHCPKAPZTITRNZFYIUZNBNP WFEVNMWEXGTTUBWGYREGRNSVTFUHAJCASVYEJVACWSLVVFJQICWNTUXWPIOLNREH HYSBXWVBHXMIAWSHUWRXQTAXJPFTHWVPMLPDFBMIMDRXUAENSYUHTTUVASXWYNHV OAAMDFWDRJNZRQSEWXDWNLLTFQLBMHVWXEXRZOMHRFWWNXJWTDEWBYPFFBSVPOVD HTNJTADNMURDIKQLIPTEPNAWUZRNQBANWOYNANHLITCBLBRFARRIAURJXQWDNGQH JSTNRRIRVWGEOLHFHVYSXPIRLFIAHYC

3. And this one:

CEJUQDBFZBCRZGPMACSVGOAWTKOTTEPVAQBJYBCNFCGEAADCJJROVJMYUHQOUMID WXWKDQRJXVOOKOVZEUGFIUNERKLYJPIZJTKRKPVHVKACCRENEHYPWKRVYQBJUARP HEZOPKPKGFRSQFFHWVCBGJHQXYNAYYPPYDSULKVFSDUHDOJGECDIXTBNRVFDLGYJ BTCOGHPKVCMRMZXRJBHXRISBPMMRXSKLPEISUEWABJAWPWSNBJBWIYKTXJIJNBFZ NQNTYFDBLYSHVFJBLSFKOZJHFWDJRQRUBCDHOHWDWIMTQWZMJRUKFXYOHQNIFHEW UDEDCBZFIRANKLYGZADRLOZESVKBQDSIQJQRUFZEDQKFPQXMSRIYBFIGFCNKIUFZ GQGXDYGIOYZRJXFJYBAOPCSOPRIMMIKSCGGRMMKTAEOKYMZRYNRULPQNTWCJUGRD LZHDFRAPUKSMRKNKNFHBVCAONJEUZWQKW