Fondamenti di Informatica I
a.a. 2007/2008

Relazione sul Progetto:
Codice per la Decifratura di Vignère con Metodo Statistico

Autore: Matteo Ruggero Ronchi
Matr. 006715

1. Introduzione

Il progetto è basato interamente sull'algoritmo di cifratura di Vigenère, così chiamato in onore di colui che lo portò a forma definitiva. Questo metodo, ideato nel XVI secolo, è tutt'ora utilizzato come metodo di cifratura in numerosi software. La forza della cifratura di Vigenère consiste nell'utilizzare non solo uno, ma fino a ventisei diversi alfabeti cifranti per crittare un solo messaggio. Pertanto, la decifratura di un codice di Vignère rappresenta un'ottima applicazione delle tecniche di programmazione studiate durante il corso di Fondamenti di Informatica I. In particolare, nel codice sviluppato per questo progetto sono stati utilizzati i seguenti elementi:

- funzioni, prototipi e chiamate
- chiamate da linea di comando in ambiente console linux (int argc, char *argv[])
- array di stringhe (puntatore a puntatore a caratteri)
- scrittura e lettura su e da matrici
- puntatori a file
- funzioni di libreria (fopen, fclose, fseek, ftell, fprintf, strlen, strcpy)
- allocazione dinamica di memoria (calloc)
- istruzione break

Il programma può essere lanciato da linea di comando e richiede in ingresso quattro stringhe di testo:

- il nome del file che contiene il testo in chiaro da cifrare: `chiaro`
- il nome del file nel quale scrivere il testo strippato: `strippato`
- il nome del file nel quale scrivere il testo cifrato: `cifrato`
- la chiave di cifratura: `chiave`

A esempio, dopo la compilazione eseguita con l'istruzione:

```
$ gcc intmain.c -o main
```

per eseguire il programma, la linea di comando dovrà essere nella forma:

```
$ ./main  chiaro strippato cifrato chiave
```

Al termine della routine verrà creato un file (`decifrato`) in cui sarà possibile visionare l'esito della decrittatura eseguita con la chiave che è stata ricavata dal programma mediante l'analisi statistica del testo cifrato.

2. La Cifratura di Vignère

Il primo passo consiste nella stesura della tavola di Vigenère. Si tratta di un normale alfabeto chiaro di 26 lettere seguito da 26 alfabeti cifranti, ciascuno spostato a sinistra di una lettera rispetto a quello precedente. Perciò ogni riga rappresenta un alfabeto cifrante con uno spostamento (di Cesare) corrispondente al numero della riga stessa:

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

La cifratura di Vigenere comporta che per ciascuna lettera del messaggio da cifrare si utilizzi una riga diversa della tavola, cioè un diverso alfabeto cifrante. La sequenza con cui si susseguono gli alfabeti cifranti è data dalla parola, o frase, chiave.

Dunque, per cifrare un messaggio, si sostituisce ogni lettera del testo in chiaro con quella che si trova nella tabella di Vigenère all'intersezione tra la colonna che inizia con il carattere del testo da cifrare con la riga corrispondente alla lettera della chiave.

Ad esempio, volendo cifrare la frase "let's meet at midnight" con la chiave "example", per prima cosa riscriviremo il messaggio senza punteggiatura e spazi, e poi scriveremo, in corrispondenza di ogni lettera, la chiave con cui cifrarla:

l e t s m e e t a t m i d n i g h t
ex am p l e e x am p l ee xa m

a questo punto è chiaro che per cifrare la 'l' dovremo usare l'alfabeto cifrante che inizia per 'e'; per cifrare la 'e' quello che inizia per 'x', e così via.

Scorrendo la tabella sulla colonna 'l' fino all'intersezione con la riga 'e' troviamo la lettera 'p', che costituirà la prima lettera del messaggio cifrato. Proseguendo così fino alla fine del testo in chiaro, il testo cifrato finale sarà:

p b t e b p i x x t y x o r m d h f

Poichè la chiave è formata da sette lettere, il messaggio sarà cifrato usando sette alfabeti differenti, passando da uno all'altro nello stesso ordine in cui si susseguono le lettere nella chiave. Dopo la settima riga - quella dell'ultima 'e' della chiave - si "torna" alla riga della 'e' - la prima della chiave - e il ciclo si ripete.

La cifratura di Vigenère appartiene alla classe dei metodi polialfabetici, poichè impiega più alfabeti cifranti per messaggio. Ed è proprio questa caratteristica il suo punto di forza: e' infatti inattaccabile con un semplice studio delle frequenze. Lettere uguali del testo in chiaro possono infatti essere rappresentate da simboli differenti nel crittogramma. Come vedremo, è però ancora possible applicare metodi statistici per la decifratura, sia pure leggermente più sofisticati.

3. Il Metodo Statistico per la Decifratura di Vignère

Prima cosa da verificare è se il testo è stato realmete cifrato con il metodo di Vigenère, oppure se è stato usato un metodo monoalfabetico (ovvero se la chiave è di una sola lettera, o se è composta da più lettere). Per questo si utilizza il *Friedman Test*[1] che si basa sulla determinazione di un parametro noto come *Indice di Coincidenza* (indicato con l'acronimo IC), che rappresenta la probabilità che due lettere casualmente selezionate in un testo risultino identiche.

[1] Friedman, Milton. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance". Journal of the American Statistical Association 32 (200): 675–701.

In dettaglio, sia *n* il numero totale di caratteri in un testo, n_1 il numero di a, n_2 il numero di b, e così via. Possiamo esprimere la probabilità che le due lettere selezionate siano entrambi delle 'a' come:

$$P(2a) = \frac{\binom{n_1}{2}}{\binom{n}{2}}$$

dove con $\binom{a}{b}$ viene indicato il coefficiente binomiale tra a e b, ovvero il rapporto tra tutti i casi favorevoli ed i casi possibili.

La probabilità che le due lettere selezionate siano entrambe delle 'b' è dunque:

$$P(2b) = \frac{\binom{n_2}{2}}{\binom{n}{2}}$$

e così per tutte le altre lettere.

In definitiva, la probabilità che due lettere casualmente selezionate in un testo siano identiche è la somma di tutte queste probabilità, ovvero:

$$IC = P(2a) + P(2b) + \ldots + P(2z)$$

dove, per ogni generica lettera:

$$\binom{n_i}{2} = \frac{n_i\,!}{2! \times \left(n_i - 2\right)\,!} = \frac{n_i \times \left(n_i - 1\right)}{2}$$

Dal momento che

$$\frac{\binom{n_i}{2}}{\binom{n}{2}} = \frac{\frac{n_i \times \left(n_i - 1\right)}{2}}{\frac{n \times (n-1)}{2}} = \frac{n_i \times \left(n_i - 1\right)}{n \times \left(n - 1\right)}$$

si ottiene

$$IC = \sum_{i=1}^{26} \frac{n_i \times (n_i - 1)}{n \times (n-1)} \approx \sum_{i=1}^{26} \left(\frac{n_i}{n} \right)^2$$

dove l'approssimazione risulta valida dal momento che generalmente $n_i , n >> 1$

Dopo aver calcolato IC, è importante studiarne il valore: per lingue quali l'inglese o l'italiano con le relative frequenze delle varie lettere, IC assume all'incirca il valore 0.065. In un linguaggio in cui ogni lettera ha la stessa frequenza (1/26), IC assume invece approssimativamente il valore 0.038.

Poichè per ogni cifratura monoalfabetica la distribuzione delle frequenze è invariante (cioè cambiano la lettere più frequenti ma non le frequenze con cui compaiono), un valore di IC vicino a 0.065 indicherà che il messaggio è stato cifrato utilizzando un metodo di cifratura monoalfabetica. Un valore numerico di IC vicino a 0.038 indicherà invece che la cifratura deriva probabilmente da un algoritmo polialfabetico.

Dopo aver effettuato lo studio del valore di IC è dunque possibile applicare l'algoritmo di decrittazione. Il primo passo è quello di determinare la lunghezza della chiave di cifratura in modo che sia successivamente possibile passare allo studio delle frequenze. Per riuscirvi si sfrutta nuovamente l'Indice di Coincidenza.

Nella routine per il calcolo della lunghezza viene infatti calcolato IC per tutte le possibili sottostringhe in cui può essere suddiviso il messaggio cifrato per una data lunghezza della chiave, che può essere al massimo di "MAX_LUNGH_KEY" caratteri. Il valore di IC di ogni sottostringa viene quindi memorizzato in un vettore il cui indice rappresenta la quantità di sottostringhe in cui il testo è stato suddiviso. Il maggior scostamento dal valore di IC calcolato sul testo cifrato iniziale indicherà la lunghezza più probabile della chiave di cifratura, che verrà memorizzata in una variabile apposita.

Una volta determinata la lunghezza della chiave, il testo cifrato viene suddiviso in tante sottostringhe quante sono le lettere costituenti la chiave stessa, ognuna delle quali è stata cifrata, come visto sopra, dallo stesso alfabeto cifrante. Questo implica che in ogni sottostringa è stato usato un solo alfabeto cifrante, e cioè che la cifratura è di natura monoalfabetica e dunque attaccabile con una analisi delle frequenze.

Per ogni sottostringa, la lettera più frequente corrisponderà con ogni probabilità alla lettera più frequente in ogni sottostringa del testo in chiaro (cioè alla lettera ‘e’, assumendo l’utilizzo della lingua inglese). Trovato il carattere più frequente, sarà sufficiente leggere dalla tabella di Vigenère con quale lettera la 'e' dovrebbe essere stata composta affinchè si ottenesse la letterà cifrata più frequente. Ripetendo questa procedura per tutte le sottostringhe generate dal testo cifrato sarà possible ottenere tutte le lettere della chiave di cifratura.

Ad esempio, nel caso in cui in una sottostringa cifrata la lettera più frequente sia la 'n', potremo ragionevolmente supporre che essa corrisponda alla 'e' del messaggio originale in chiaro. Per ottenere la lettera tramite la quale la 'e' è stata cifrata in 'n', sarà sufficente leggere sulla tabella di Vigenère lungo la riga 'e' fino ad incontrare la lettera 'n'. Dal momento che la colonna in cui appare la 'n' è iniziata dalla lettera 'j' si ottiene che questa è la lettera corrispondente della chiave di cifratura.

Come ultimo passo occorre semplicemente invertire il metodo di Vigenère per ricavare dal testo cifrato il messaggio originale utilizzando ciclicamante le lettere della chiave di cifratura. Il testo cifrato ci appariva così:

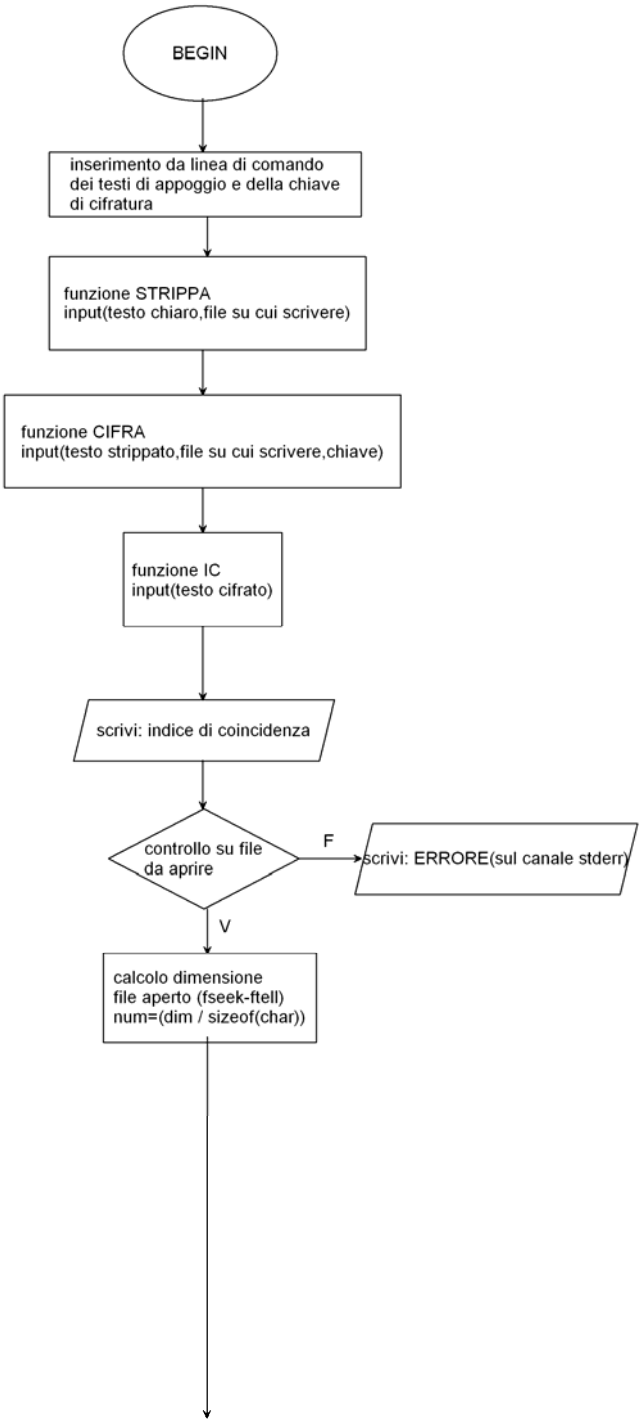
p b t e b p i x x t y x o r m d h f

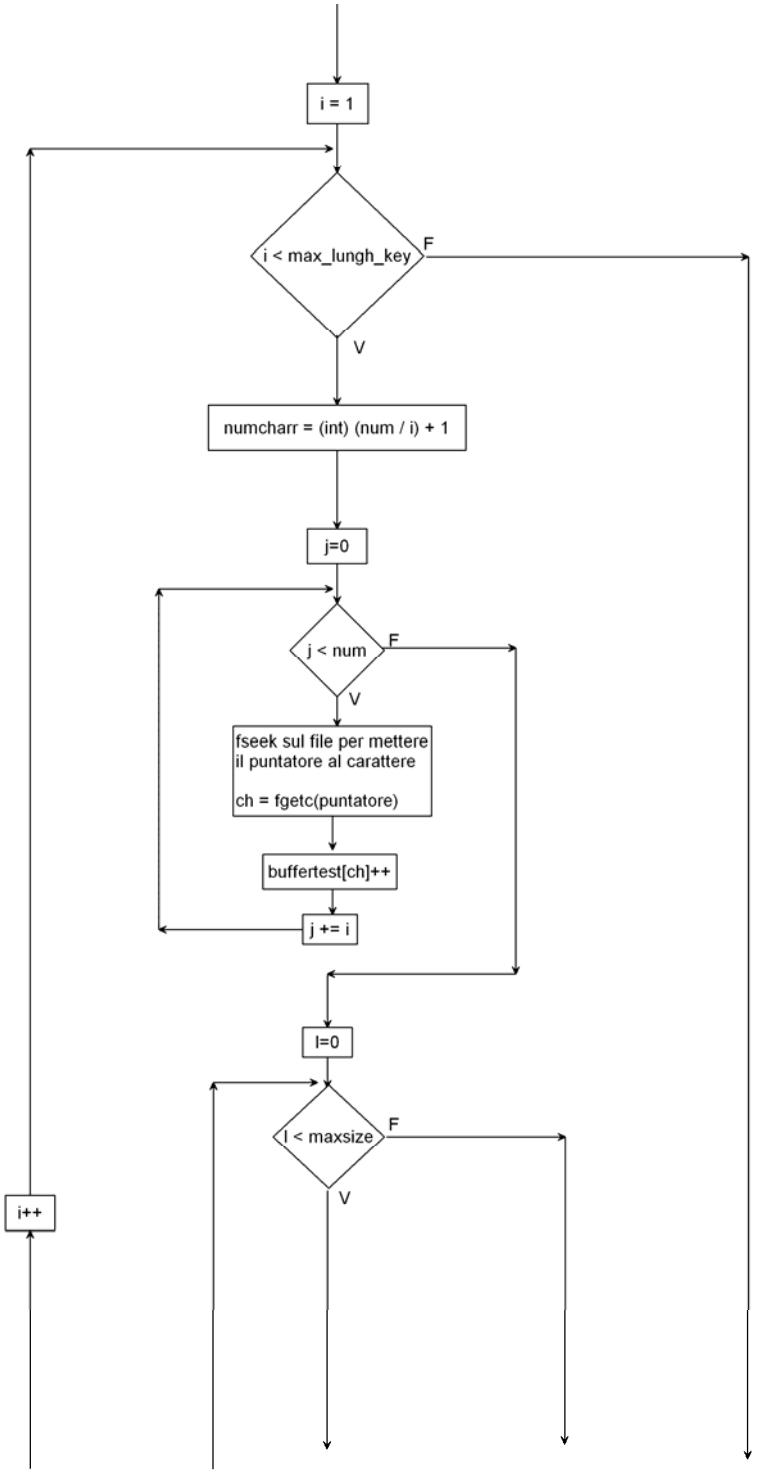
e abbiamo scoperto la chiave essere “example”; sapendo che la prima lettera del testo cifrato è la 'p' e la prima della chiave è 'e' ricerchiamo nella colonna della 'e' lungo quale riga appare la 'p': scopriamo che è la riga iniziata dalla 'l', ecco la prima lettera del testo originale. Scorrendo tutta la chiave e ripetendola ciclicamente, applicando questa routine per ogni carattere del testo cifrato, possiamo ricostruire interamente il messaggio originale.

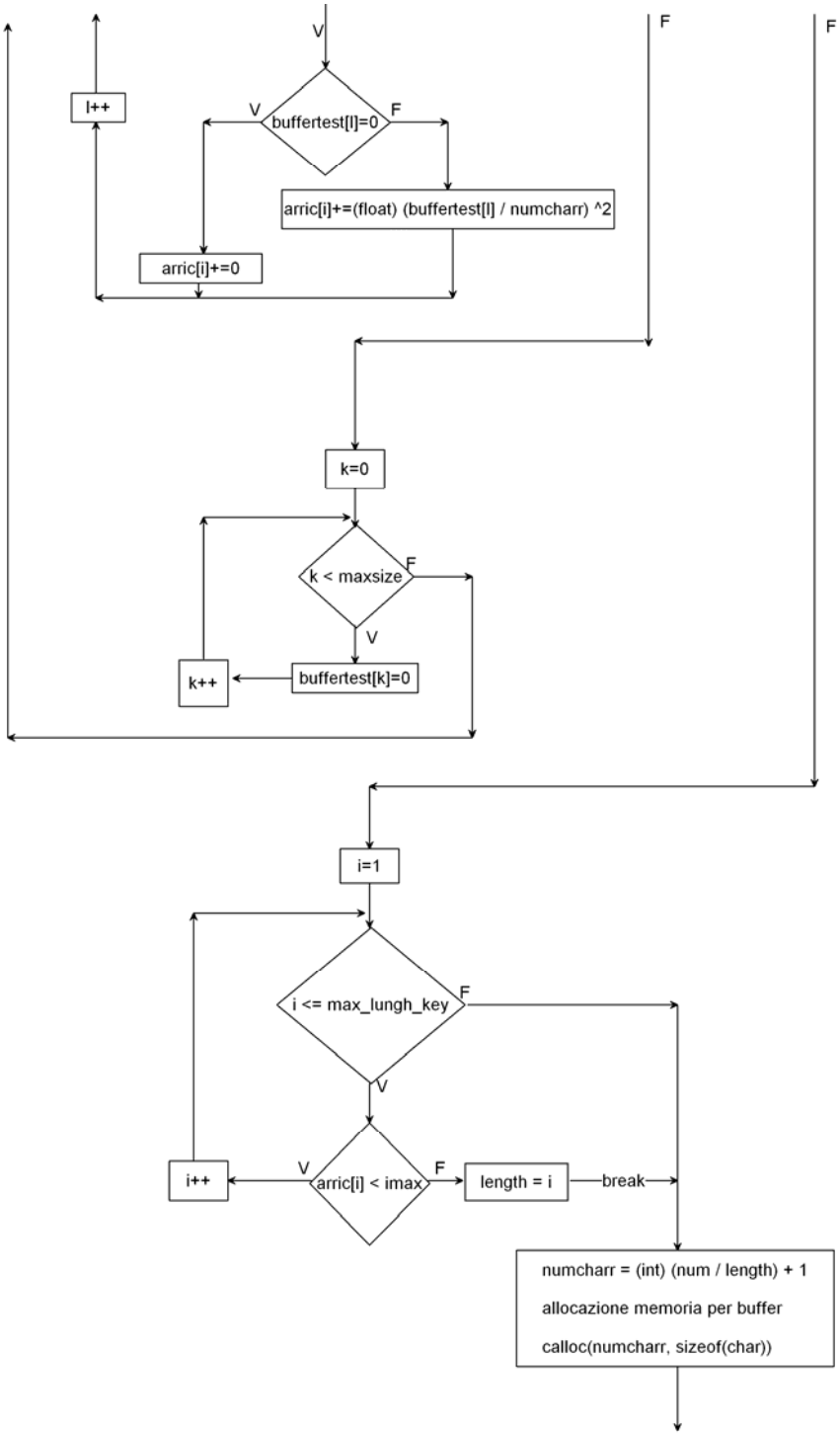
p b t e b p i x x t y x o r m d h f
e x a mp l e e xa mp l e e x am
l e t sm e e t at mi d n i g ht

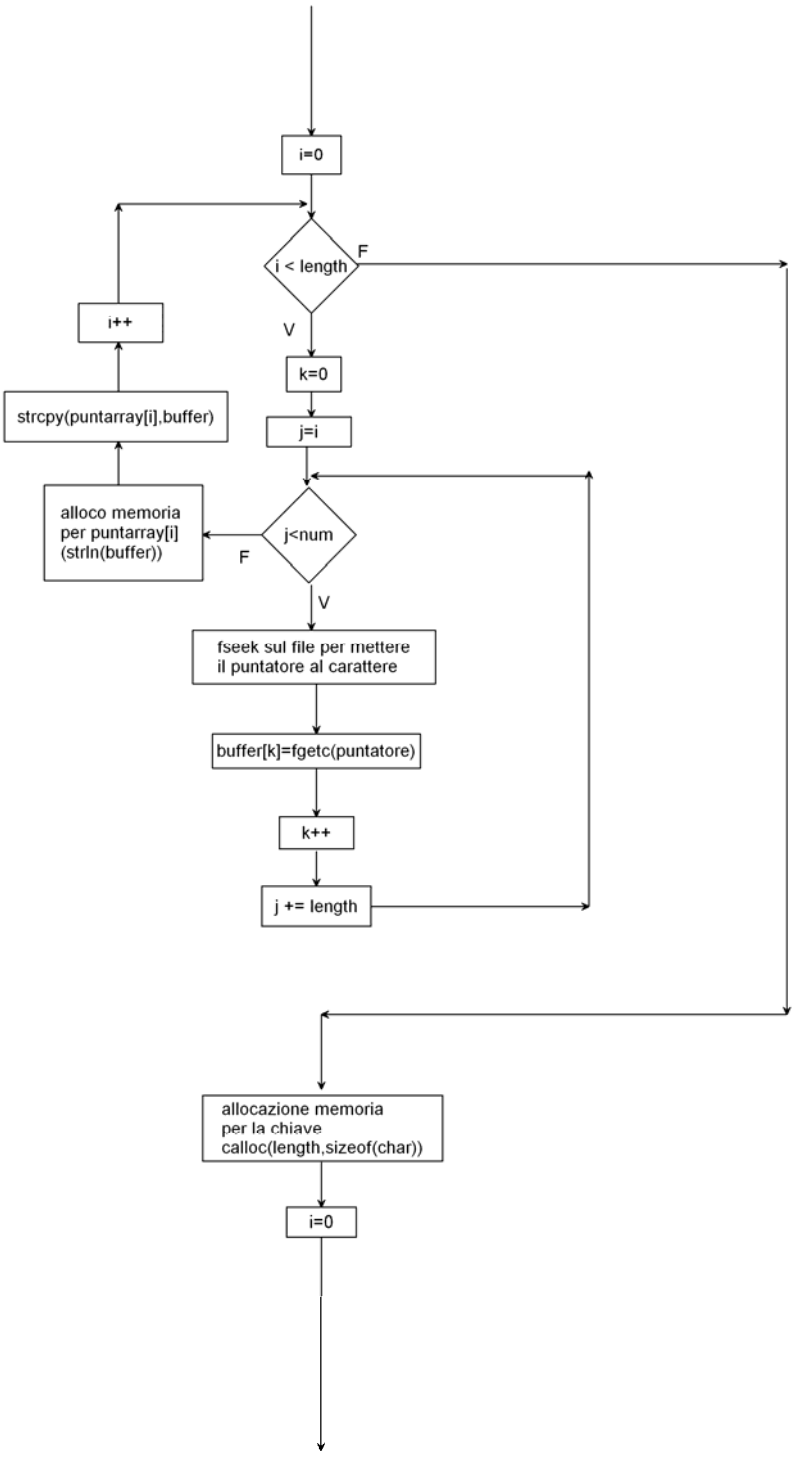
4. Diagrammi di Flusso

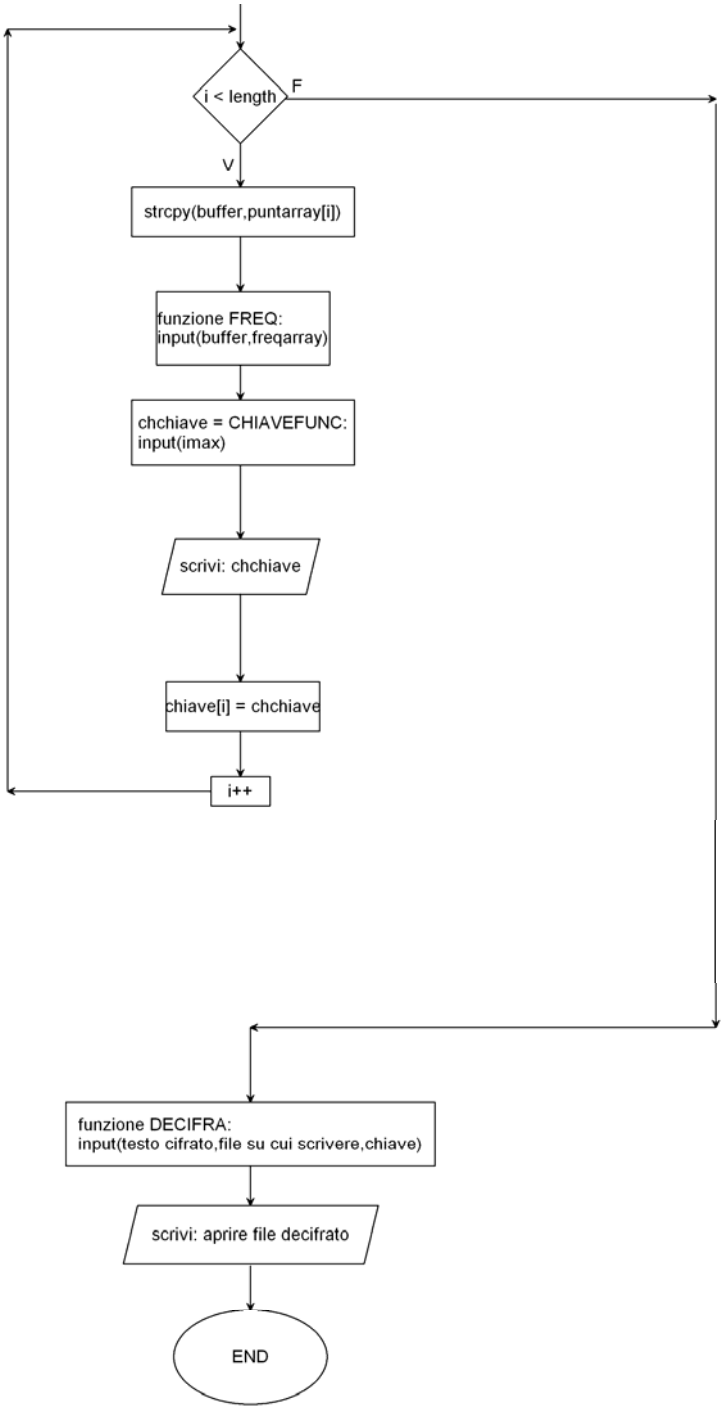
Int main




i = 1
i < max_lungh_key
F
V
numcharr = (int) (num / i) + 1
j=0
j < num
F
V
fseek sul file per mettere
il puntatore al carattere
ch = fgetc(puntatore)
buffertest[ch]++
j += i
l=0
l < maxsize
F
V
i++

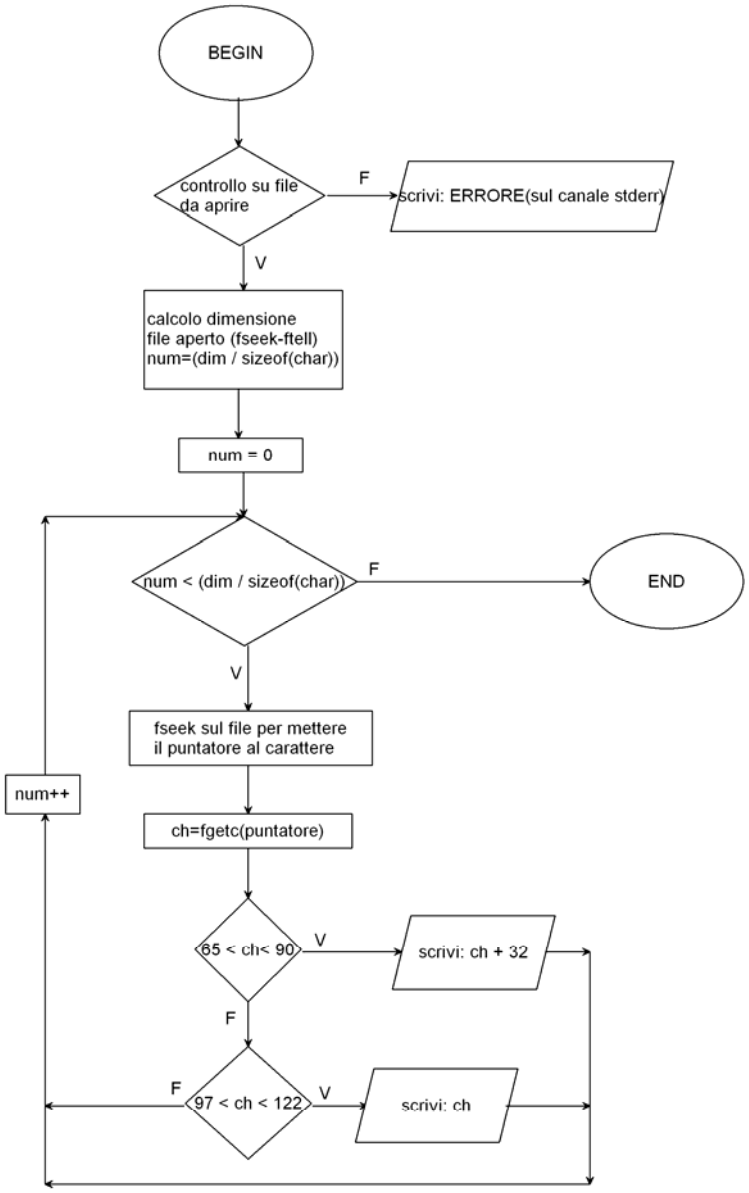

V
F
F
l++
V
buffertest[l]=0
F
arric[i]+=(float) (buffertest[l] / numcharr) ^2
arric[i]+=0
k=0
k < maxsize
F
V
k++
buffertest[k]=0
i=1
i <= max_lungh_key
F
V
i++
V
arric[i] < imax
F
length = i
break
numcharr = (int) (num / length) + 1
allocazione memoria per buffer
calloc(numcharr, sizeof(char))


i=0
i < length
F
V
i++
k=0
strcpy(puntarray[i],buffer)
j=i
alloco memoria
per puntarray[i]
(strln(buffer))
j<num
F
V
fseek sul file per mettere
il puntatore al carattere
buffer[k]=fgetc(puntatore)
k++
j += length
allocazione memoria
per la chiave
calloc(length,sizeof(char))
i=0


i < length
F
V
strcpy(buffer,puntarray[i])
funzione FREQ:
input(buffer,freqarray)
chchiave = CHIAVEFUNC:
input(imax)
scrivi: chchiave
chiave[i] = chchiave
i++
funzione DECIFRA:
input(testo cifrato,file su cui scrivere,chiave)
scrivi: aprire file decifrato
END

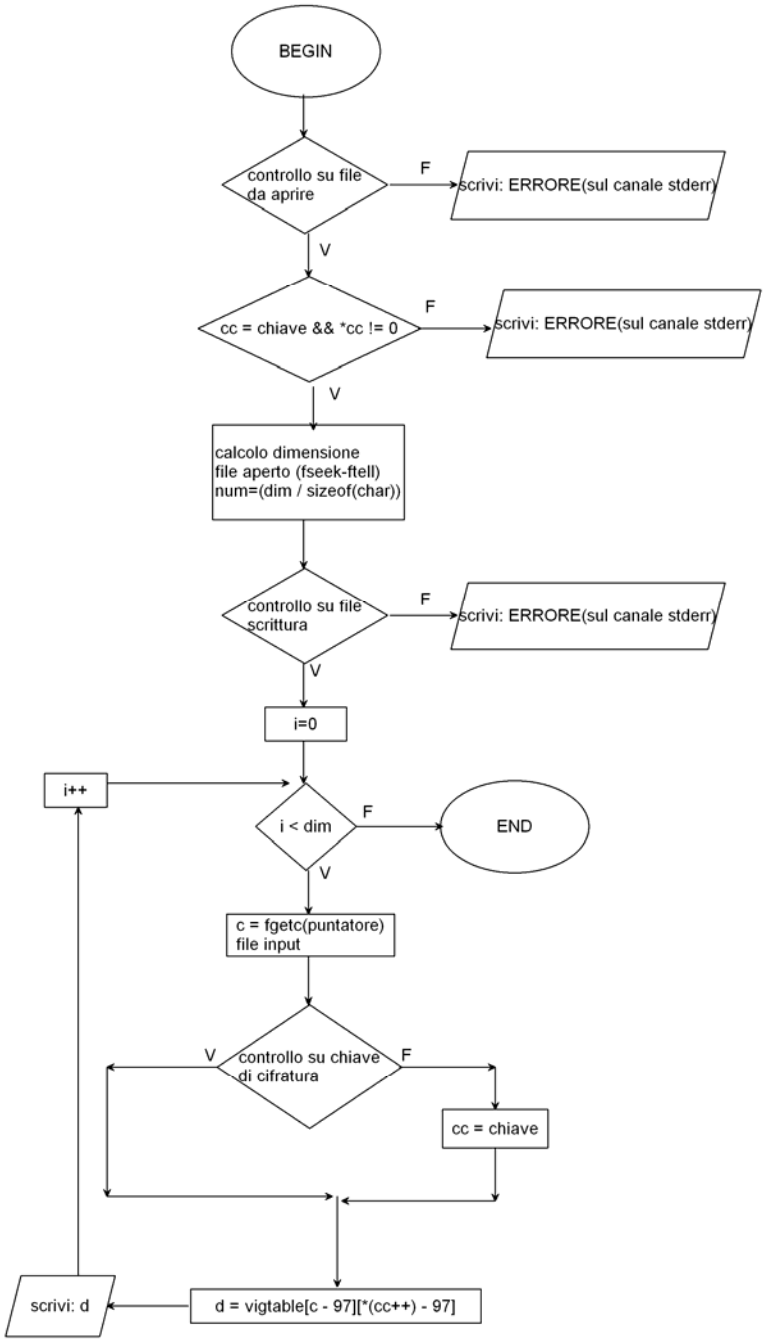
Funzione strippa

Prende come input il puntatore al file, con testo in chiaro (lingua inglese), di cui si desidera effettuare la cifratura, e il puntatore al file su cui si desidera copiare il testo risultante dalla operazione di strippatura. Questa funzione viene utilizzata per ripulire il testo da cifrare, da tutta la punteggiatura ed i caratteri speciali, cosi' da rendere in seguito possibile la cifratura utilizzando l'algoritmo di Vigenère. Stampa quindi il testo strippato sul file indicato dal puntatore datole in ingresso.



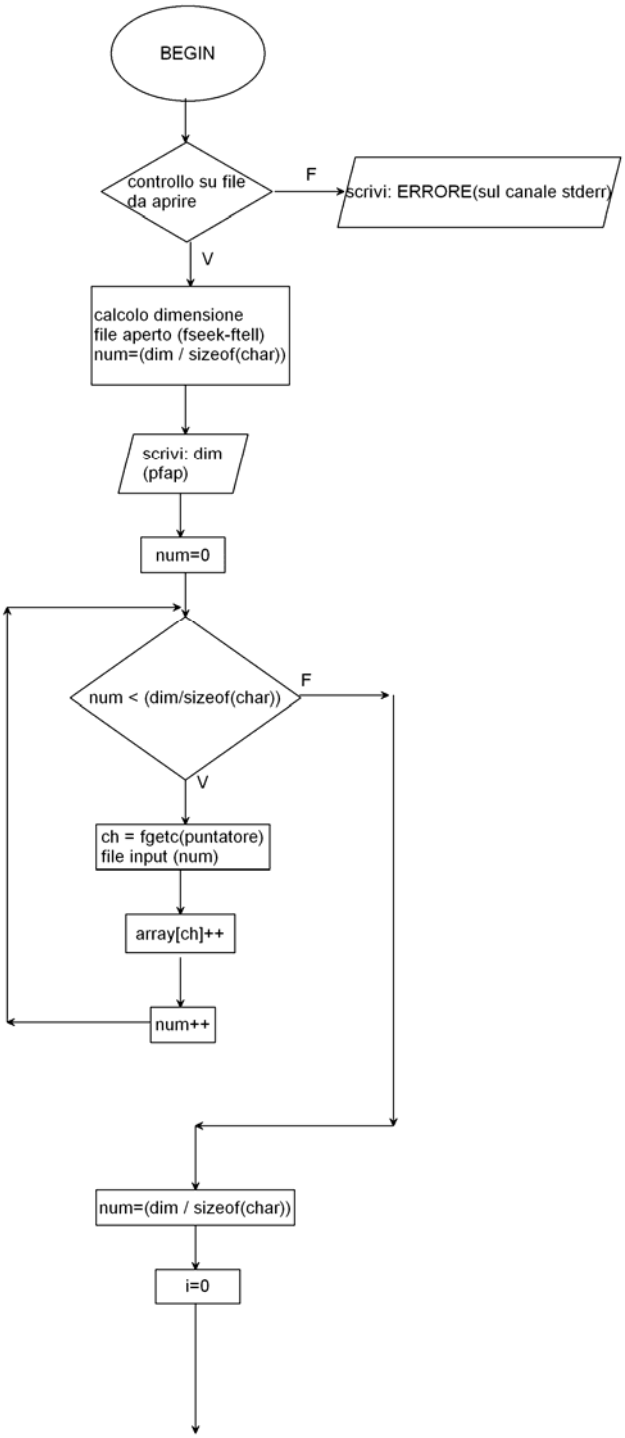
Funzione cifra

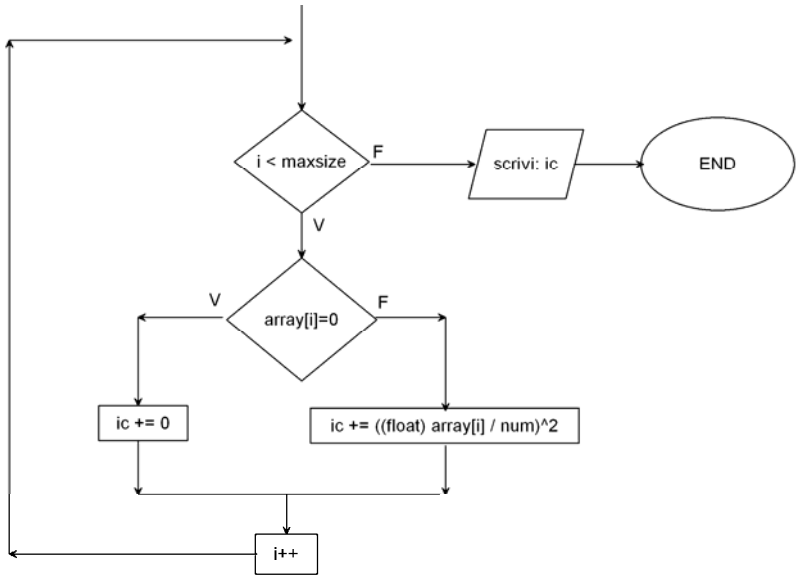
Prende come input il puntatore al file, strippato mediante la funzione 'strippa', di cui si desidera effettuare la cifratura, il puntatore al file su cui si desidera copiare il testo risultante dalla operazione di cifratura, e la chiave mediante la quale bisogna implementare il metodo di Vigenère. La funzione, dopo aver generato la tabella di Vigenère, ne utilizza l'algoritmo per effettuare la trasposizione in codice. Esegue la cifratura di ogni carattere nel testo con una diversa lettera della chiave, la quale viene poi ripetuta il numero necessario di volte affinchè tutto il testo risulti cifrato. Infine stampa il risultato dell'algoritmo sul file indicato dal puntatore passatole in ingresso.



Funzione Indice di Coincidenza

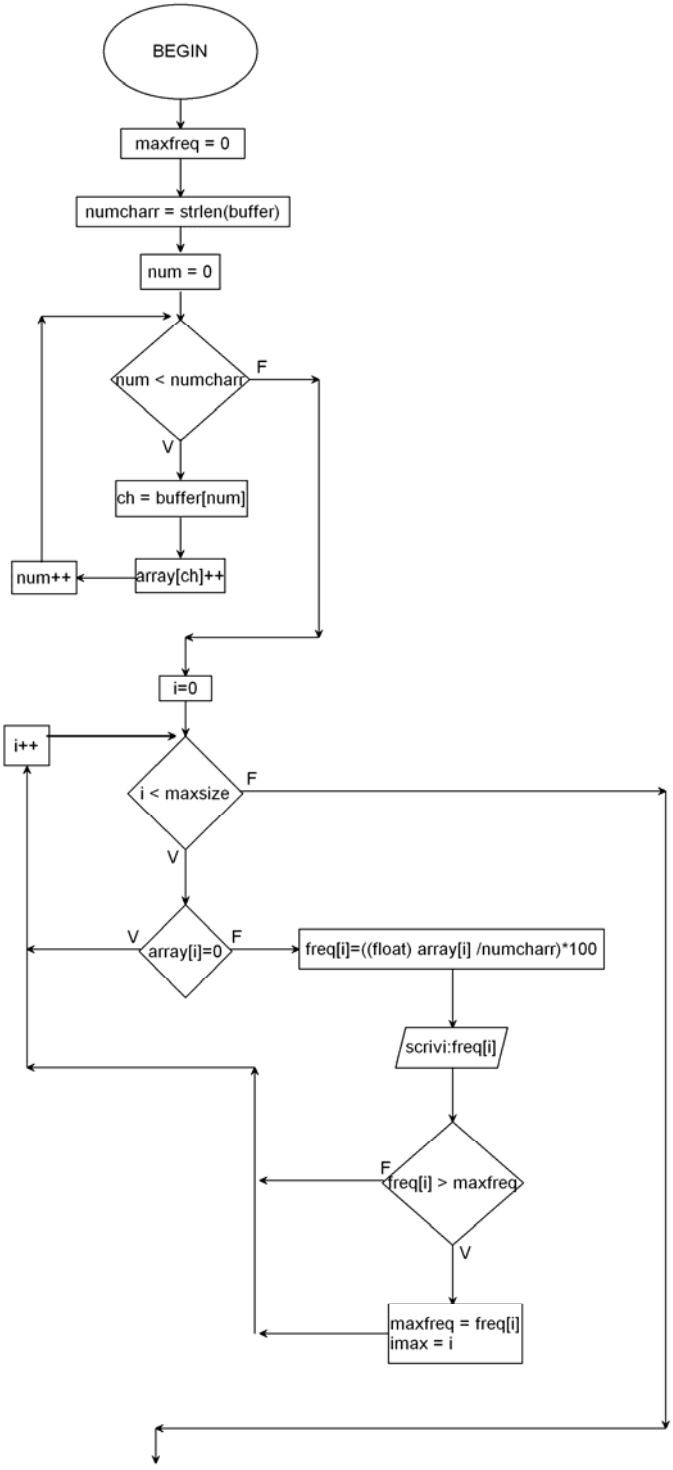
Questa funzione prende in ingresso il puntatore al file cifrato e restituisce il valore (in doppia precisione) dell'Indice di Coincidenza relativo al testo.

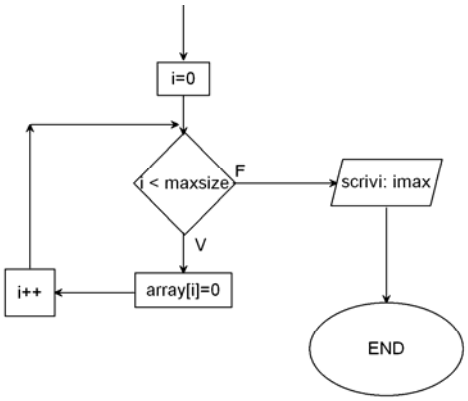



i < maxsize
F
scrivi: ic
END
V
V
array[i]=0
F
ic += 0
ic += ((float) array[i] / num)^2
i++

Funzione freq

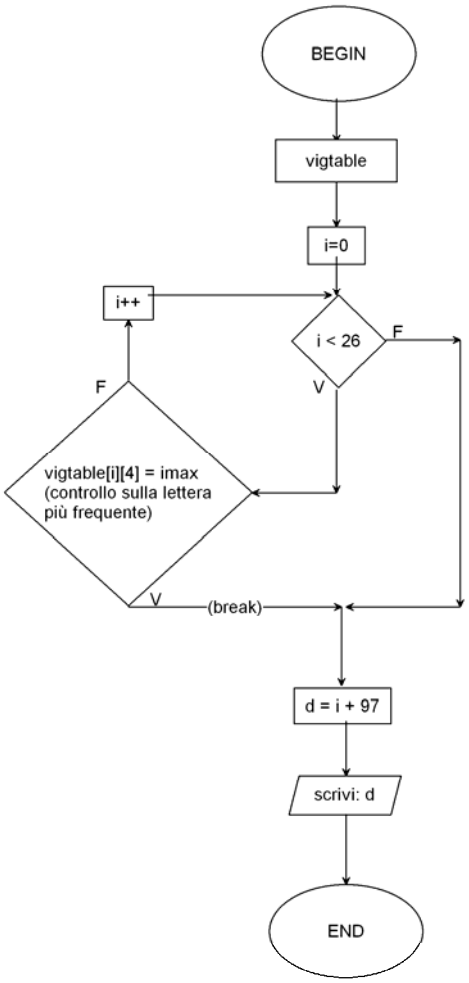
La funzione prende come input un puntatore a caratteri (buffer) sul quale sono stati copiati temporaneamente tutti i caratteri di una delle sottostringe del testo cifrato, e un array di float sul quale memorizza le frequenze relative alla sottostringa data in ingresso. La routine studia la frequenza con cui si ripetono i caratteri presenti nella sottostringa in esame e ne memorizza il piu' frequente, del quale viene restituito il valore ASCII corrispondente.




i=0
i < maxsize
F
V
scrivi: imax
array[i]=0
i++
END

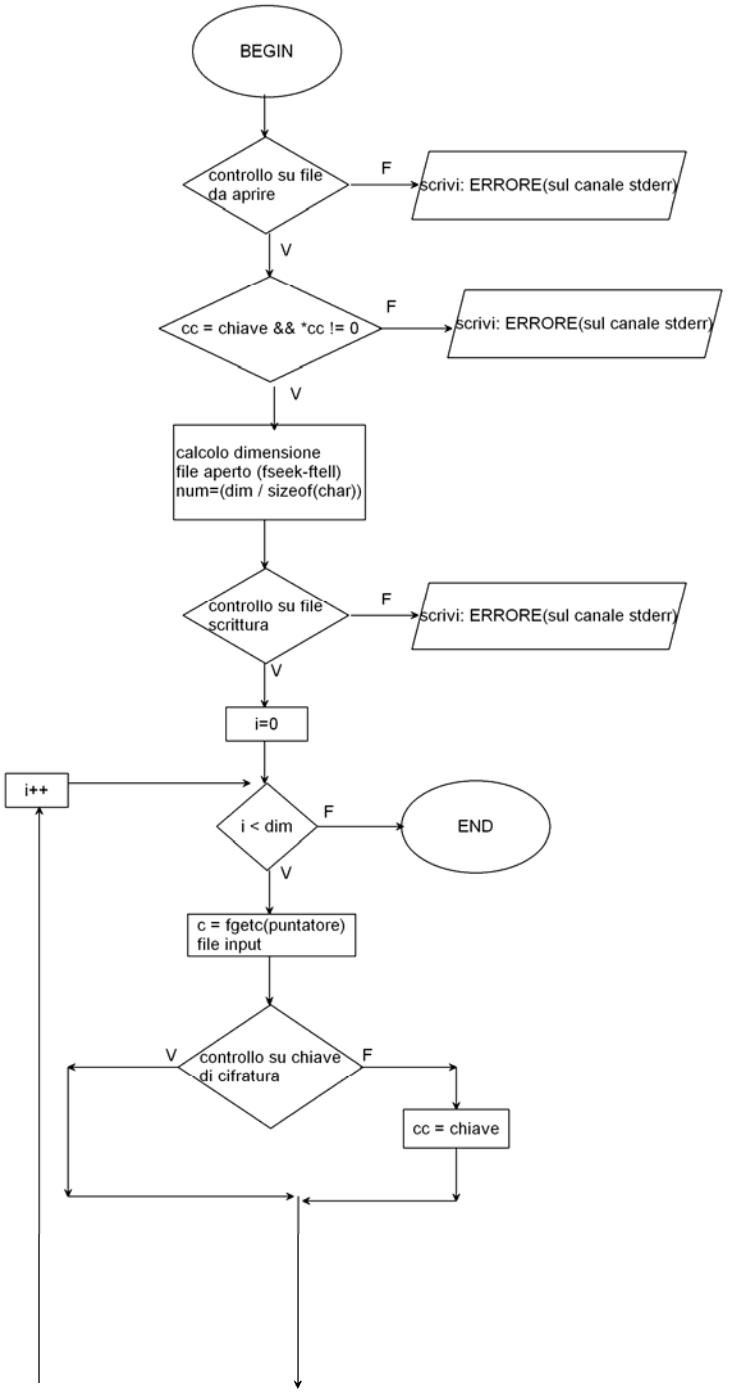
Funzione chiavefunc

L'input e' costituito dal valore ASCII del carattere piu' frequente della sottostringa presa in esame dalla funzione `freq`. Supponendo che nel testo in chiaro (in lingua inglese) questo corrisponda alla lettera 'e' la funzione ricava mediante la tabella di Vigenère la lettera con la quale la 'e' debba essere stata cifrata per ottenere il carattere `IMAX`. Questa e' dunque una delle lettere costituenti la chiave e rappresenta l'output della funzione.



Funzione decifra

Prende come input il puntatore al file cifrato, il puntatore al file su cui si desidera scrivere il testo decrittato mediante la routine della funzione, e la chiave di cifratura ricavata dalla esecuzione delle precedenti parti del programma, mediante la quale bisognerà implementare la decrittatura. La funzione si appoggia alla tabella di Vigenère, ed esegue un confronto fra ogni singolo carettere del testo cifrato e l'elemento della tabella con indice di colonna la lettera corrispondente alla chiave di cifratura (ricavata) e di riga una variabile contatore. Quando il confronto dà esito positivo, il carattere corrispondente all'indice di riga viene scritto sul file decrittato (puntato dal puntatore a file passato come argomento); la routine e' ripetuta finche' tutto il testo cifrato e' stato scorso.



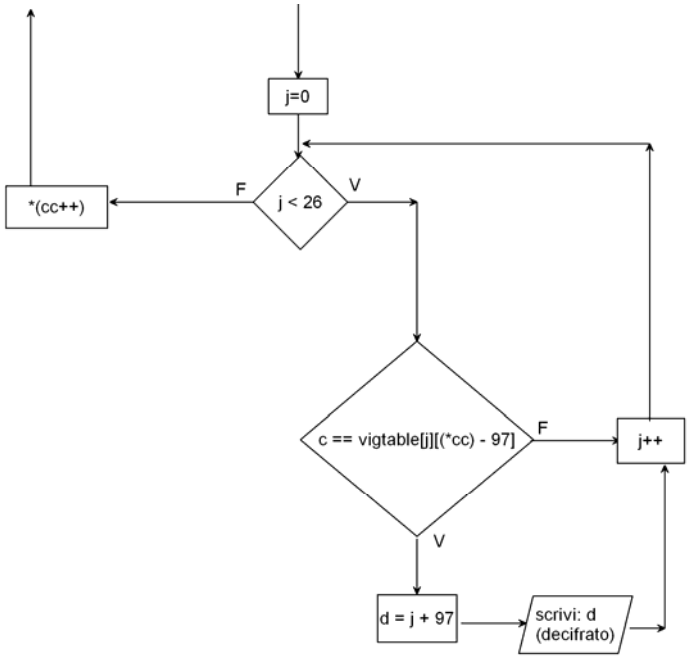

j=0
j < 26
F
V
*(cc++)
c == vigtable[j][(*cc) - 97]
F
j++
V
d = j + 97
scrivi: d
(decifrato)

Tavola di Vigenère

Questo è l'algoritmo utilizzato per implementare la tabella di Vigenère nelle funzione richiamate dal programma.