Definition

A genomic sequence is a sequence over the alphabet of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) nucleotides.

The primary structure of DNA can be represented as a sequence over the alphabet of nucleotides: the purines A (adenine) and G (guanine), and the pyrimidines C (cytosine) and T (thymine).

The primary structure of RNA can also be represented as a sequence over the alphabet of nucleotides, where the pyrimidine T (thymine) is replaced by U (uracil).

The primary structure of a protein can be represented as a sequence over the alphabet of amino acids or residues: the hydrophobic or water-insoluble A (alanine), C (cysteine), I (isoleucine), L (leucine), M (methionine), F (phenylalanine), and V (valine), and the hydrophilic or water-soluble R (arginine), N (asparagine), D (aspartate), E (glutamate), Q (glutamine), G (glycine), H (histidine), K (lysine), P (proline), S (serine), T (threonine), W (tryptophan), and Y (tyrosine).

Example

The DNA sequences AAAGGAGGTGGTCCA and TGGACCACCTCCTTT, which are common to most organisms, are complementary.

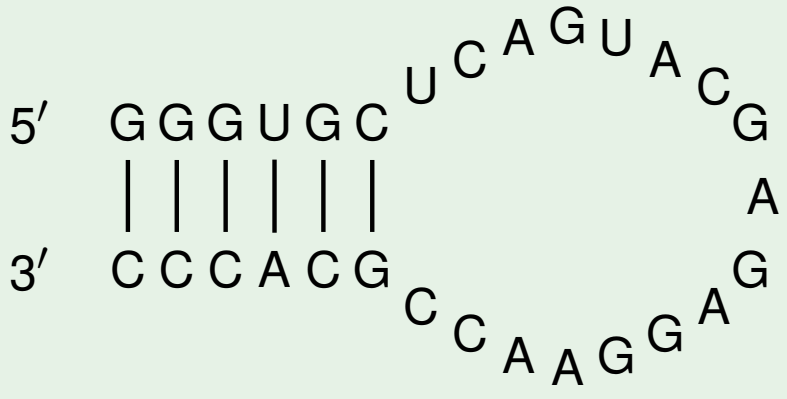
```
5′-AAAGGAGGTGGTCCA-3′
3′-TGGACCACCTCCTTT-5′
```

Example

The large subunit ribosomal RNA sequence

```
5′-GGGUGCUCAGUACGAGAGGAACCGCACCC-3′
```

folds back on itself forming the sarcin/ricin loop, a highly conserved form of RNA secondary structure across different species.



Example

The following protein sequence, `2bop`, corresponds to a DNA binding protein from a bovine virus, the Bovine papillomavirus.

```
SCFALISGTANQVKCYRFRVKKNHRHRYENCTTTWFTVADNGAERQGQAQILI
TFGSPSQRQDFLKHVPLPPGMNISGFTASLDF
```

Definition

DNA consists of two strands of complementary nucleotides that fold in space in the shape of a double helix, each of whose two ends has the $3'$ end of one DNA strand and the $5'$ end of the other DNA strand.

Reading one such sequence in the $5'$ to $3'$ direction, the complementary sequence will be read in the reversed, $3'$ to $5'$ direction, and corresponding nucleotides in the two sequences will be complementary.

Example

The following DNA sequence is also shown reversed, complemented, and reverse complemented.

```
5′-AAAGGAGGTGGTCCA-3′
3′-ACCTGGTGGAGGAAA-5′
5′-TTTCCTCCACCAGGT-3′
3′-TGGACCACCTCCTTT-5′
```

Algorithm

The reverse of a DNA sequence can be obtained by traversing the DNA sequence from the $3'$ to the $5'$ end, and the reverse complement can be obtained by replacing the nucleotides by their complementary nucleotides during the traversal.

A DNA sequence S of length n is reversed by putting, for each i with $1 \leqslant i \leqslant n$, the ith nucleotide in position $n - i + 1$ of the reverse sequence, and it is reverse complemented by also replacing each nucleotide with the complementary one.

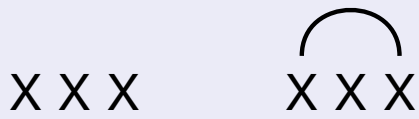
Algorithm

function reverse_complement(S)
    $n \leftarrow \text{length}(S)$
    for $i \leftarrow 1$ **to** n **do**
        if $S[i] = \texttt{A}$ **then**
            $R[n-i+1] \leftarrow \texttt{T}$
        else if $S[i] = \texttt{C}$ **then**
            $R[n-i+1] \leftarrow \texttt{G}$
        else if $S[i] = \texttt{G}$ **then**
            $R[n-i+1] \leftarrow \texttt{C}$
        else if $S[i] = \texttt{T}$ **then**
            $R[n-i+1] \leftarrow \texttt{A}$
    return R

Definition

An interesting problem consists in counting the number $R(n)$ of possible RNA secondary structures of length n.

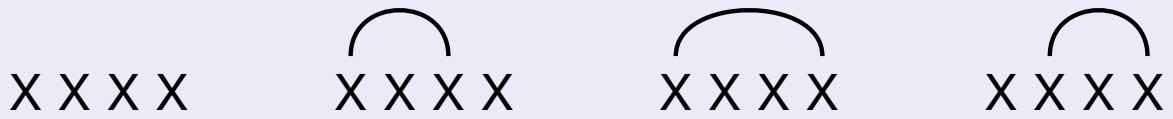
While there is only one possible RNA secondary structure of length 0 (the empty sequence) and only one possible RNA secondary structure of length 1 or 2, there are two possible RNA secondary structures of length 3,



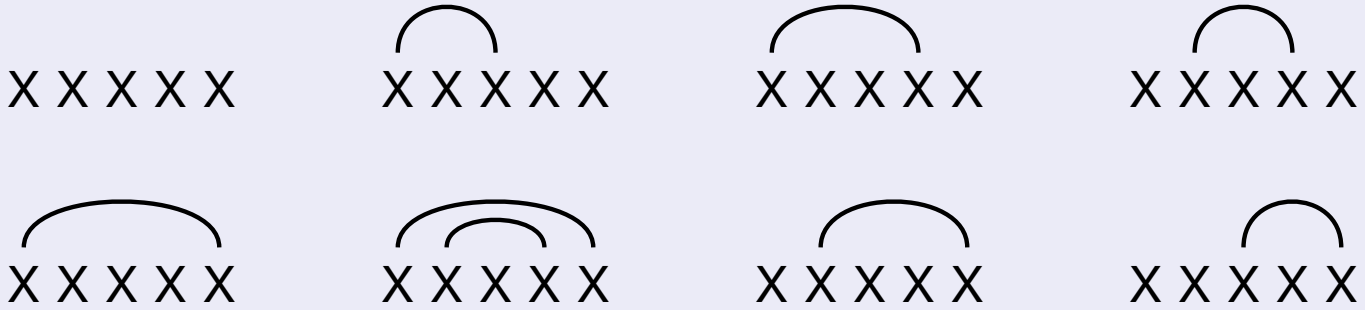
where X stands for an A, C, G, U base along the RNA sequence, in $5'$ to $3'$ order.

Definition

Further, there are four possible RNA secondary structures of length 4,



and eight possible RNA secondary structures of length 5,



Definition

In general, in a sequence of length $n+1$, the base at position $n+1$ is either not paired or it is paired with the base at position j, where $1 \leqslant j \leqslant n-1$.

In the latter case, the bases at positions 1 through $j-1$ can form any of the $R(j-1)$ possible secondary structures of length $j-1$, and the bases in positions $j+1$ through n can also form any of the $R(n-j)$ possible secondary structures of length $n-j$.

Therefore, $R(0) = R(1) = R(2) = 1$ and, for $n \geqslant 2$,

$$R(n+1) = R(n) + \sum_{j=1}^{n-1} R(j-1)R(n-j)$$

Algorithm

This gives an algorithm for counting the number $R(n)$ of RNA secondary structures of length n.

The computation of $R(n+1)$ requires the values of $R(j-1)$ and $R(n-j)$ for each $j = 1, \ldots, n-1$, that is, it requires each of the values $R(0), R(1), \ldots, R(n-1)$.

These values are computed in that order and stored in a vector, so that they are already available whenever needed during the computation of $R(n)$.

Algorithm

function count(n)
    $R[0] \leftarrow R[1] \leftarrow R[2] \leftarrow 1$
    for $i = 2$ **to** $n - 1$ **do**
        $R[i+1] \leftarrow R[i]$
        for $j = 1$ **to** $i - 1$ **do**
            $R[i+1] \leftarrow R[i+1] + R[j-1] \cdot R[i-j]$
    return $R[n]$

Definition

There is only one possible RNA secondary structure of length 0 (the empty sequence) and only one possible RNA secondary structure of length 1.

In general, in a sequence from position i to position j, the base at position j is either not paired or it is paired with the base at position k, where $i \leqslant k < j$.

In the latter case, the bases at positions i through $k-1$ can form any of the $R(i, k-1)$ possible secondary structures of length $k-i$, and the bases in positions $k+1$ through $j-1$ can also form any of the $R(k+1, j-1)$ possible secondary structures of length $j-k-1$.

Therefore, $R(i,j) = 1$ for $1 \leqslant i = j+1 \leqslant n$ and for $1 \leqslant i = j \leqslant n$ and, for $1 \leqslant i < j \leqslant n$,

$$R(i,j) = R(i,j-1) + \sum_{k=i}^{j-1} R(i,k-1)R(k+1,j-1)$$

Algorithm

	X	X	X	X	X	X	X	X
X	1	1	2	4	8	17	37	82
X		1	1	2	4	8	17	37
X			1	1	2	4	8	17
X				1	1	2	4	8
X					1	1	2	4
X						1	1	2
X							1	1
X								1

Algorithm

```
function count(n)
    for i = 1 to n do
        for j = 1 to n do
            R[i,j] ← 0
        R[i,i] ← 1
    for j = 2 to n do
        for i = 1 to j − 1 do
            R[i,j] ← R[i,j − 1]
            for k = i to j − 1 do
                R[i,j] ← R[i,j] + R[i,k − 1] ∗ R[k + 1,j − 1]
    return R[1,n]
```

Example

The following large subunit ribosomal RNA sequence, of length 29,

```
GGGUGCUCAGUACGAGAGGAACCGCACCC
```

has 8,622,571,758 possible secondary structures, only 789,564 of which are indeed possible for this RNA sequence.

The remaining 8,621,782,194 secondary structures involve base pairs other than AU or CG.

Definition

All the DNA sequences with n nucleotides can be generated by taking each of the DNA sequences with $n-1$ nucleotides in turn and then extending them with one more nucleotide.

Example

The four DNA sequences of length 1,

A C G T

can each be extended in four different ways to give sequences of length 2.

AA	AC	AG	AT
CA	CC	CG	CT
GA	GC	GG	GT
TA	TC	TG	TT

Example

Each of these DNA sequences of length 2 can in turn be extended in four different ways to give sequences of length 3.

AAA	AAC	AAG	AAT	ACA	ACC	ACG	ACT
AGA	AGC	AGG	AGT	ATA	ATC	ATG	ATT
CAA	CAC	CAG	CAT	CCA	CCC	CCG	CCT
CGA	CGC	CGG	CGT	CTA	CTC	CTG	CTT
GAA	GAC	GAG	GAT	GCA	GCC	GCG	GCT
GGA	GGC	GGG	GGT	GTA	GTC	GTG	GTT
TAA	TAC	TAG	TAT	TCA	TCC	TCG	TCT
TGA	TGC	TGG	TGT	TTA	TTC	TTG	TTT

Algorithm

Thus, the DNA sequences of length 1 are just the elements of the alphabet $\Sigma = \{A, C, G, T\}$, and the DNA sequences of length $n > 1$ are the result of extending the DNA sequences of length $n - 1$ with an element of Σ.

This gives an algorithm for generating all DNA sequences of length $n \geqslant 1$.

Algorithm

function words(n, Σ)
    if $n = 1$ **then**
        $L \leftarrow \Sigma$
    else
        $S \leftarrow \text{words}(n-1, \Sigma)$
        $L \leftarrow \emptyset$
        for each word w of S **do**
            for each element s of Σ **do**
                $w' \leftarrow \text{concat}(w, s)$
                $L \leftarrow L \cup \{w'\}$
    return L