Computing GC Content

Identifying Unknown DNA Quickly

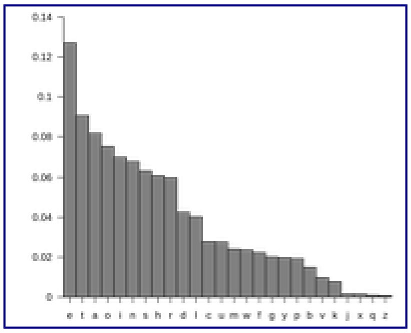

Figure 1. The table above was computed from a large number of English words and shows for any letter the frequency with which it appears in those words. These frequencies can be used to reliably identify a piece of English text and differentiate it from that of another language. Taken from http://en.wikipedia.org/wiki/File:English_letter_frequency_(frequency).svg.

A quick method used by early computer software to determine the language of a given piece of text was to analyze the frequency with which each letter appeared in the text. This strategy was used because each language tends to exhibit its own letter frequencies, and as long as the text under consideration is long enough, software will correctly recognize the language quickly and with a very low error rate. See Figure 1 for a table compiling English letter frequencies.

You may ask: what in the world does this linguistic problem have to do with biology? Although two members of the same species will have different genomes, they still share the vast percentage of their DNA; notably, 99.9% of the 3.2 billion base pairs in a human genome are common to almost all humans (i.e., excluding people having major genetic defects). For this reason, biologists will speak of *the* human genome, meaning an average-case genome derived from a collection of individuals. Such an average case genome can be assembled for any species, a challenge that we will soon discuss.

The biological analog of identifying unknown text arises when researchers encounter a molecule of DNA from an unknown species. Because of the base pairing relations of the two DNA strands, cytosine and guanine will always appear in equal amounts in a double-stranded DNA molecule. Thus, to analyze the symbol frequencies of DNA for comparison against a database, we compute the molecule's GC-content, or the percentage of its bases that are *either* cytosine or guanine.

In practice, the GC-content of most eukaryotic genomes hovers around 50%. However, because genomes are so long, we may be able to distinguish species based on very small discrepancies in GC-content; furthermore, most prokaryotes have a GC-content significantly higher than 50%, so that GC-content can be used to quickly differentiate many prokaryotes and eukaryotes by using relatively small DNA samples.

Example

The GC-content of a DNA string is given by the percentage of symbols in the string that are 'C' or 'G'. For example, the GC-content of "AGCTATAG" is 37.5%. Note that the reverse complement of any DNA string has the same GC-content.

DNA strings must be labeled when they are consolidated into a database. A commonly used method of string labeling is called FASTA format. In this format, the string is introduced by a line that begins with '>', followed by some labeling information. Subsequent lines contain the string itself; the first line to begin with '>' indicates the label of the next string.

In Rosalind's implementation, a string in FASTA format will be labeled by the ID "Rosalind_xxxx", where "xxxx" denotes a four-digit code between 0000 and 9999.

Given: At most 10 DNA strings in FASTA format (of length at most 1 kbp each).

Return: The ID of the string having the highest GC-content, followed by the GC-content of that string. Rosalind allows for a default error of 0.001 in all decimal answers unless otherwise stated; please see the note on absolute error below.

Sample Dataset

```
>Rosalind_6404
CCTGCGGAAGATCGGCACTAGAATAGCCAGAACCGTTTCTCTGAGGCTTCCGGCCTTCCC
TCCCACTAATAATTCTGAGG
>Rosalind_5959
CCATCGGTAGCGCATCCTTAGTCCAATTAAGTCCCTATCCAGGCGCTCCGCCGAAGGTCT
ATATCCATTTGTCAGCAGACACGC
>Rosalind_0808
CCACCCTCGTGGTATGGCTAGGCATTCAGGAACCGGAGAACGCTTCAGACCAGCCCGGAC
TGGGAACCTGCGGGCAGTAGGTGGAAT
```

Sample Output

```
Rosalind_0808
60.919540
```

Counting Point Mutations

Evolution as a Sequence of Mistakes

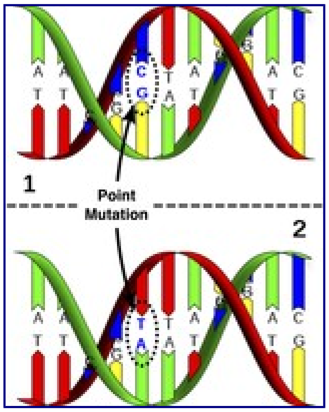

Figure 1. A point mutation in DNA changing a C-G pair to an A-T pair.

A mutation is simply a mistake that occurs during the creation or copying of a nucleic acid, in particular DNA. Because nucleic acids are vital to cellular functions, mutations tend to cause a ripple effect throughout the cell. Although mutations are technically mistakes, a very rare mutation may equip the cell with a beneficial attribute. In fact, the macro effects of evolution are attributable by the accumulated result of beneficial microscopic mutations over many generations.

The simplest and most common type of nucleic acid mutation is a point mutation, which replaces one base with another at a single nucleotide. In the case of DNA, a point mutation must change the complementary base accordingly; see Figure 1.

Two DNA strands taken from different organism or species genomes are homologous if they share a recent ancestor; thus, counting the number of bases at which homologous strands differ provides us with the minimum number of point mutations that could have occurred on the evolutionary path between the two strands.

We are interested in minimizing the number of (point) mutations separating two species because of the biological principle of parsimony, which demands that evolutionary histories should be as simply explained as possible.

Example

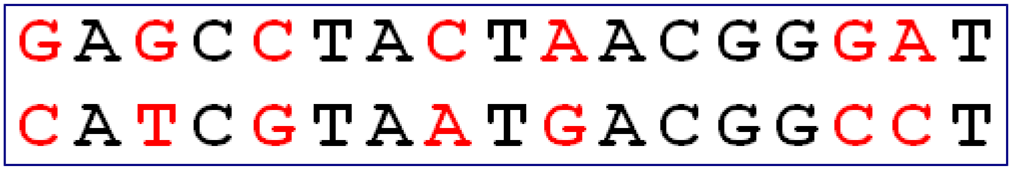

Figure 2. The Hamming distance between these two strings is 7. Mismatched symbols are colored red.

Given two strings *s* and *t* of equal length, the Hamming distance between *s* and *t*, denoted *d*H(*s*,*t*), is the number of corresponding symbols that differ in *s* and *t*. See Figure 2.

Given: Two DNA strings *s* and *t* of equal length (not exceeding 1 kbp).

Return: The Hamming distance *d*H(*s*,*t*).

Sample Dataset

```
GAGCCTACTAACGGGAT
CATCGTAATGACGGCCT
```

Sample Output

```
7
```