WEEK 3: THE CLASSIFICATION TASK

Summary:

* Our goal is to discover a genetic mutation that can be used as a diagnostic test for cancer.
* A perfect test will always result in a positive result for an individual with cancer and a negative result for an individual who does not have cancer.
* We evaluated the effectiveness of two genetic mutations as possible diagnostic tests:
  + *RNF43\_GRCh37\_17:56435161-56435161\_Frame-Shift-Del\_DEL\_C-C--*
* *RPL22\_GRCh37\_1:6257785-6257785\_Frame-Shift-Del\_DEL\_T-T—*

Conclusions:

* The mutation *RNF43\_GRCh37\_17:56435161-56435161\_Frame-Shift-Del\_DEL\_C-C--* (which yields 11+105 correct classifications and 45+1 incorrect classifications) is more useful for diagnosis than the mutation *RPL22\_GRCh37\_1:6257785-6257785\_Frame-Shift-Del\_DEL\_T-T--* (which yields 9+104 correct classifications and 47+2 incorrect classifications).
* Across the ~1,400 genetic mutations provided in the data file, the best test will result from using the genetic mutation DOCK3\_GRCh37\_3:51417604-51417604\_Frame-Shift- Del\_DEL\_C-C--, which yields a TP-FP value of 12 (the maximum TP-FP value for *any* mutation) and a %TP-%FP value of 7.4%.
* There is no single genetic mutation that can be used as a highly accurate diagnostic test.

Results:

A **True Positive** occurs when an individual with cancer is correctly diagnosed as being in the **C** class (i.e., the individual gets a positive test result). For example, in our data set we have 56 samples in the **C** class. If we use the mutation *RNF43\_GRCh37\_17:56435161-56435161\_Frame-Shift-Del\_DEL\_C-C--* as a diagnostic test, we will correctly diagnose 11 of the **C** individuals (because they have the mutation), resulting in 11 *True Positives*. Approximately 20% of the individuals with cancer will receive a positive result if this mutation of the RNF43 gene is used as a diagnostic test.

A **False Positive** occurs when a cancer-free individual is misdiagnosed as being in the **C** class (i.e., the individual gets a positive test result). In this case, a cancer-free individual is given a positive test result. For example, in our data set we have 106 samples in the **NC**class. If we use the mutation *RNF43\_GRCh37\_17:56435161-56435161\_Frame-Shift-Del\_DEL\_C-C--* as a diagnostic test, we will incorrectly diagnose 1 of the **NC** individuals (because they have the mutation), resulting in 1 *False Positive*. Approximately 1% of the cancer-free individuals will receive a positive result if this mutation of the RNF43 gene is used as a diagnostic test.

A **False Negative** occurs when an individual with cancer is misdiagnosed as being in the **NC** class. In this case, an individual with cancer is given a negative test result. For example, in our data set we have 56 samples in the **C**class. If we use the mutation *RNF43\_GRCh37\_17:56435161-56435161\_Frame-Shift-Del\_DEL\_C-C--* as a diagnostic test, we will correctly diagnose 11 of the **C** individuals (because they have the mutation).  However, we will misdiagnose 56-11=45 individuals in the **C** class, resulting in 45 *False Negatives*. Approximately 80% of the individuals with cancer will receive a negative result if this mutation of the RNF43 gene is used as a diagnostic test.

A **True Negative** occurs when a cancer-free individual is correctly diagnosed as being in the **NC** class. For example, in our data set we have 106 samples in the **NC** class. If we use the mutation *RNF43\_GRCh37\_17:56435161-56435161\_Frame-Shift-Del\_DEL\_C-C--* as a diagnostic test, we will misdiagnose 1 of the **NC** individuals (because they have the mutation).   However, we will correctly diagnose 106-1=105 individuals in the **NC** class, resulting in 105 *True Negatives*. Approximately 99% of the cancer-free individuals will receive a negative result if this mutation of the RNF43 gene is used as a diagnostic test.

Below are two confusion matrices, summarizing the effectiveness of the KRAS and RPL22 mutations that were studied.

*RNF43\_GRCh37\_17:56435161-56435161\_Frame-Shift-Del\_DEL\_C-C--*

*RPL22\_GRCh37\_1:6257785-6257785\_Frame-Shift-Del\_DEL\_T-T--*

Table 1. Confusion matrix for the use of the genetic mutation RNF43\_GRCh37\_17:56435161-56435161\_Frame-Shift-Del\_DEL\_C-C-- as a diagnostic test for cancer.

|  |  |  |
| --- | --- | --- |
|  | P (actual **C**) | N (actual **NC**) |
| P (predicted **C**) | 11 TPs | 1 FPs |
| N (predicted **NC**) | 45 FNs | 105 TNs |

Table 2. Confusion matrix for the use of the genetic mutation *RPL22\_GRCh37\_1:6257785-6257785\_Frame-Shift-Del\_DEL\_T-T--* as a diagnostic test for cancer.

|  |  |  |
| --- | --- | --- |
|  | P (actual **C**) | N (actual **NC**) |
| P (predicted **C**) | 9 TPs | 2 FPs |
| N (predicted **NC**) | 47 FNs | 104 TNs |