Генетический смысл данных

Информация об организме может быть закодирована последовательностью (*гаплотипом*) на алфавите из четырех символов: `A`, `C`, `G`, `T`, (*нуклеотидов*) например,

```
ATCCCTAT
```

У некоторых организмов, в том числе у людей, гаплотипа *два*: одна копия от papa, вторая от maman. Пара гаплотипов называется *генотипом*. Например:

```
ATCCCTAT
ATCCGTAT
```

Видно, что в пятой позиции гаплотипы отличаются нуклеотидом. Это называется *Single Nucleotide Polymorphism* (SNP).

Последовательность нуклеотидов кодирует *ген*. Гены определяют признаки индивида (например, цвет глаз) — *фенотип*. Гены могут иметь варианты (*аллели*), например в 5-й позиции могла стоять комбинация `(C;C)` или `(G;G)`, что могло бы соответствовать разным фенотипам (например, разному цвету глаз).

Замечание. Часто аллель используют в значении «нуклеотид в конкретной позиции конкретного гаплотипа», а генотип — в значении «конкретная комбинация аллелей».

Некоторые аллели встречаются чаще других в популяции. К примеру, в приведенной последовательности пусть `C` встречается чаще `G`. Тогда `C` называется major, а `G` — minor аллелем. Информацию о генотипе поэтому удобно хранить в виде одной последовательности на алфавите 0, 1, 2, что соответствует major или minor аллелю в обоих гаплотипах, или разнице в типе. Например:

```
011000110
001100010
---------
021200210
```

Иногда добавляют «3» для индикации отсутствия данных по позиции.