METHOD TO COMPRESS GENOMIC DATA WITHOUT LOSS

ZULFIKAR MOINUDDIN AHMED

Our key is that a 64-bit computer double precision floating point has 52 bits. Now uniqueness of individuals is determined not based on all 3 billion DBA base pairs in germ cells or twice so in somatic cells but by 0.1% of this, i.e. roughly 6 *million* snips (SNPs) of mostly binary possibilities.

The simplest scheme is thus to take probabilities for minor alleles across all snips and order them with 2 significant digits. For each p_j of distinct minor allele probability, we can find A_j snip locations where human genome can vary. Then we can just order the data as blocks of A_j bits and p_j. If we do not store various annotations for each of the snips, keeping them far away from the data, how much is the compression? It's 52 snips/double roughly. For 6.5 million snips per human chromosome, we could have 100 probabilities $p_1, \ldots, p_{100}$ and 125,000 doubles. A double is 16 bytes, so this data fits into 2 megabytes; rough estimate gets us 40-50 Mb for the entire genetic uniqueness of individual human beings, for obviously this gives us information about uniqueness.

1. Substantial Issue is Collection of 40-50 Mb Per Human

The binary compression is not substantial novelty but important to understand. What we need are measurement of the 40-50 Mb of data from every person in the planet and we want to make that easy and cheap for them. It is only with exact data for every person that we can produce products and services to cater to every person on Earth in a rational manner for substantial benefit and increased *Life Satisfaction* for them. That is where the real business is from my viewpoint.

I do not know how to engineer chips that can quickly determine the allele for all parts of the genome. Then the measurements can be passed and handled in a large database. Per-person 40-50 Mb is actually quite small and this can then lead to large class of products and services because this amount fits in easily in smartphones or USB type devices.

2. Uniqueness Is The Key

Genetic uniqueness with 40-50 Mb of data measured per person is a transformative event in the history of the human race, because genetic uniqueness has full information about a great deal of secondary inferences, and will change all aspects of life of human beings in the future.

Date: June 14, 2021.