A note on redundancy calculus

Because genomic and cDNA libraries can present biases towards some sequences due to the method used for the generation of the inserts and/or due cloning artifacts, monitoring redundancy is a crucial measure in any large scale sequencing project. Taking into account some of the concepts defined above, EGene calculates redundancy using two different criteria:

- sequence redundancy: number of consensus sequences generated;
- base redundancy: number of novel DNA positions discovered in respect to the input sequences.

The first criterion, *sequence redundancy*, particularly appropriate for EST sequencing projects, tries to compute how many different transcript reconstructs were attained and, according to the redundancy, checking if it is worth continuing the sequencing effort on the library(ies). The second criterion, *base redundancy*, is finer grained, evaluating how many new DNA positions are being found by the sequencing project. Thus, if a set of reads present short overlaps with one another, they would be considered highly redundant according to the former criterion, but with a low base redundancy in this latter criterion. The equations below show how these two redundancy levels are calculated.

$$\text{Sequence redundancy} = 1 - \frac{c + s}{i},$$

where c, s and i represent the number of all contigs, singlets and input sequences

$$\text{Base redundancy} = 1 - \frac{\sum_{c \in contigs} size(c) + \sum_{s \in singlets} size(s)}{\sum_{i \in input} size(i)},$$

where *input* represents the set of all input sequences

The criteria above are used for both `assemble_cap3.pl` and `assemble_phrap.pl` components.