Computer Analyses of Overlapping Genes in Mycoplasma Genitalium

Yuko Osada t94092yo@sfc.keio.ac.jp
Ryo Matsushima t94403rm@sfc.keio.ac.jp
Masaru Tomita mt@sfc.keio.ac.jp

Laboratory of Bioinformatics
Department of Environmental Information
Keio University
5322 Endo, Fujisawa, 252 Japan

Gene overlapping is a phenomenon in which two adjacent genes share a portion of their coding regions. We systematically extracted overlapping genes from the genome sequences of *Mycoplasma genitalium* and *Haemophilus influenzae* produced by The Institute for Genomic Research(TIGR). We analyzed those overlapping genes with respect to their direction, length, and amino acid sequence. The following four types of overlapping genes (Figure 1) exist in the *M.genitalium* genome.

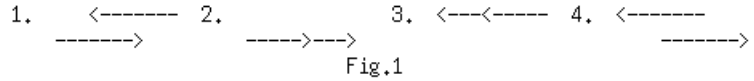

Fig.1

organism	No. of genes	No. of overlapping genes	→ ←	→ →	← ←	← →
M.genitalium	467	164	27	86	47	4
H.influenzae	1680	270	9	117	133	11

Fig.2

One of the overlapping genes, MG106, codes for formylmethionine deformylase protein. The same proteins of other species (*E.coli*, *H.influenzae* and *.T.aquatics*) were also obtained from the SWISS-PROT database. Those four sequences were aligned[1] and shown in Figure 3. We notice that all four sequences have similar N-termini except the sequence of *M.Genitalium*. Furthermore, we can find a potential start codon (marked with an arrow) at the position of start codons of the other species.

```
DEF_ECOLI  ----------------------------------MSVLQVLHIPD-ERLRKVAKPVEEV
DEF_HAEIN  ----------------------------------MTALNVLIYPD-DHLKVVCEPVTKV
DEF_THETH  ----------------------------------MVYPIRLYGD-PVLRRKARPVEDF
MG106      MLLPTPLGPVMTKILPWLFTSIVRIILTLLFLSMTFQPTKTWLVFDDNALINKPTEAV
                                  Fig.3         ↑
```

This indicates that there is a possibility of mis-identification of the start codon for the MG106 gene. This kind of errors may exist in other overlapping genes of types 2, 3 and 4. On the other hand, we can consider that type 1 overlapping genes are more reliable, because stop codons are usually not misidentified. Thus, we focused on those type 1 overlapping genes (27 pairs or 54 genes) in further analyses. Out of the 54 genes, 23 have a homologous gene in *E.coli*. Pairwise alignment with a corresponding *E.coli* gene by ClustalW show that in many sequences little or no similarity was found in the overlapping region. This means that the overlapping

[1]We used ' FASTA' mail service of Human Genome Center, Institute of Medical science, The University of Tokyo

region do not include sequences which are biologically important. Figure 4 shows a list of the overlapping genes in which the functions of the both genes are known. N-termini of both amino acid sequences are placed in the left. The regions indicated by the arrows (<===>) are overlapping with each other.

Fig.4

As we can see in those results, when one of the overlapping genes conserves the overlapping region, the other gene does not conserve the region well. This indicates that amino acid sequences of the overlapping region are biologically important in only one (or none) of the two proteins.

Thus it can be inferred that overlapping genes have emerged from two non-overlapping genes, one of the genes extending its coding region by, for example, changing its stop codon. Amino acid sequences of those extended regions would not have biological importance, but still preserved because they are harmless. Whereas some people think that overlapping genes are the results of strong evolutionary pressure to down-size genome, we conclude that it is probably not true in the case of *M.genitalium*.

Acknowledgement

This work was supported in party by a Grant-in-Aid for Scientific Research on Priority Areas 'Genome Science' from The Ministry of Education, Science, Sports and Culture in Japan.