Sequence Analysis of Short Tandem Repeats in the Genomes of *H. influenzae* and *M. genitalium*

Takanori Washio [1 2] washy@sfc.keio.ac.jp
Masahiko Wada [1 3] t93520mw@sfc.keio.ac.jp
Masaru Tomita [1 3] mt@sfc.keio.ac.jp

[1]Laboratory for Bioinformatics
[2]Graduate School of Media and Governance
[3] Department of Environmental Information
Keio University
5322 Endo, Fujisawa, 252 Japan

We systematically extracted short tandem repeats in nucleic acid and amino acid sequences from the genomes of *Haemophilis influenzae*[1] and *Mycoplasma genitalium*[2]. For each of the repeats, its sequence pattern, length, direction, location in the genome, location with respect to genes were analyzed and presented in Table 1.

The following observations have been made:

- For most of the amino acid tandem repeats, their nucleic acid sequences are also perfect repeats, implying that they have biological roles, if any, in nucleic acid level.
- All of the tandem AACC repeats (GGTT in complement) in *H. influenzae* are located in front of or at the beginning of genes coding for iron-binding proteins: HI0635, HI0712, HI1565, HI0661 and HI1385. These tandem repeats may have some role in regulation of gene expression, like the repeats in the *lic gene* in *H. influenzae* and the *opa gene* in *N. gonorrhoeae*[3][4].
- There are three trimer repeats AGT, CTT, ACT in the non-coding region between MG339 and MG340 in *M. genitalium*. Since the AGT repeats and the ACT repeats are complementary sequences, they may form a stem loop. The CTT repeats are known to form a triple strand.

We are currently investigating biological significance of each phenomenon. We also plan to extend our analysis to genomes of other procaryotes as well as eucaryotes using the entire GenBank database.

CDS	Start	Coding	Direction	Total length	Repeats	Nuc pattern	Amin pattern
HI0119	133717	cd	→	24	4	no regularity	HD
HI0216	233360	cd	→	48	4	cttaccagcgag	LTSE
HI0543	566964	cd	→	36	2	no regularity	GMG
HI0687	731229	cd	→	27	6	ttta	YLFI
HI1565	1633204	cd	→	76	19	aacc	QPTN
HI1566	1633204	cd	→	76	19	aacc	QPTN
HI0106	112703	cd	←	33	3	xccxccxax	GGL
HI0251	283098	cd	←	26	4	ctctgg	EP
HI0258	288750	cd	←	91	22	tgtc	TDRQ
HI0526	549680	cd	←	36	3	caaatttaggct	EPKE
HI0550	570798	cd	←	93	23	ttga	NQSI
HI0662	705896	cd	←	82	20	ggtt	QPTN
HI0712	760525	cd	←	148	37	ggtt	QPTN
HI1056	1122921	cd	←	129	32	actg	SQSV
HI1058	1122921	cd	←	129	32	actg	SQSV
HI0261::HI0262	291617	ncd	→::→	29	9	tta	
HI0870::HI0871	922080	ncd	→::→	62	12	ttatc	
HI1536::HI1537	1608030	ncd	→::→	71	17	aatc	
HI0352::HI0354	379520	ncd	←::→	132	33	ttga	
HI1287::HI1288	1368888	ncd	←::→	23	4	tctcg	
HI1460::HI1462	1543151	ncd	←::→	100	25	ttgc	
HI0635::HI0636	677132	ncd	←::←	84	21	ggtt	
HI1385::HI1386	1481219	ncd	←::←	66	16	aacc	
HI1717::HI1718	1788949	ncd	←::←	42	4	ccttgttcg	
MG192	227130	cd	→	35	11	agt	SS
MG309	384461	cd	←	28	4	ctatta	SN
MG338	425824	cd	←	34	11	gtt	TT
MG139::MG140	169475	ncd	→::→	50	16	agt	
MG339::MG340	429308	ncd	←::←	29	9	agt	
MG339::MG340	429967	ncd	←::←	50	16	ctt	
MG339::MG340	430015	ncd	←::←	26	8	act	
MG031::MG032	36790	ncd	←::→	19	19	a	
MG287::MG288	351452	ncd	→::←	32	10	agt	

Table 1. Short tandem repeats

Acknowledgment

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

References

[1] Robert D. Fleischmann et al., "Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd," *science*, Vol. 269, pp. 496-512, 1995.

[2] Claire M. Fraser et al., "The Minimal Gene Complement of Mycoplasma genitalium," *science*, Vol. 270, pp. 397-403, 1995.

[3] Crosa, J.H., "Genetics and molecular biology of siderophore-meddiated iron transport in bacteria," *Microbiol. Rev.*, Vol. 53, pp. 517-530, 1989.

[4] Meyer, T.F. et al., "Variation and control of protein expression in Nesseria," *Annu. Rev. Microbiol.*, Vol. 44, pp. 451-477, 1990.