Microbial genome analysis using the G-language system

Haruo Suzuki

Home · G-language Project - Documentations

http://www.g-language.org/wiki/

Home · G-language Project - Doc...

Let's start exploring the secrets of Life.
G-language Bioinformatics Environment.

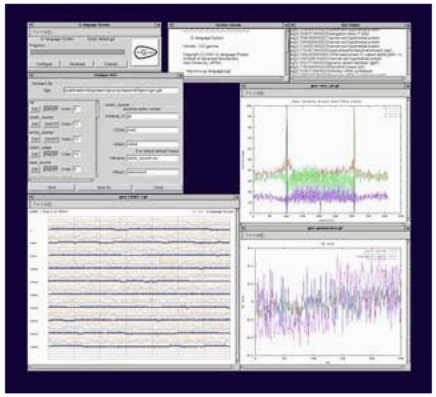
Trace: » Home

Welcome!

The G-language Genome Analysis Environment is a generic genome analysis environment aiming to:

1. Construct an integrated environment for the development of analysis software.
2. Systematically accumulate existing analysis software methodologies for analysis and their results.
3. Construct generic analysis packages that allow users to avoid redundancy in the process of analysis.

G-language Genome Analysis Environment provides a greater variety of useful genome analysis tools compared to most existing analysis software packages, and is also easily pluggable. All of its tools are accessible as Perl modules . Its Bacteria Analysis System enables users to quickly and effectively analyze bacterial genomes by simply inputting their Genbank files. Its Graphical User Interface is friendly to users who are unfamiliar with computer programming. Users can easily add programs to the G-language system as well as edit existing programs, when in need for additional analysis methods. The G-language Genome Analysis Environment will save programming time, and will guide users into a higher level of genome informatics.



MENU

- Home
- Software / Download
  - Changelogs
  - Mailing Lists
  - Development
  - Screenshots
- Documentations
  - AJAX Document Center
  - G-language Cookbook
  - Chaos Game Representation
- G-language Bookmarklet
  - Custom Bookmarklet Generator
- REST Web Service
  - REST Web Service for EMBOSS
  - REST Web Service for KEGG API
  - REST Web Service for COPASI
  - REST Web Service for E-Cell
- SOAP Web Service
- G-language Maps
  - GenomeProjector

Genes, Genomes and Genomics ©2008Global Science Books

Computational Genome Analysis Using The G-language System

Kazuharu Arakawa[§*] • **Haruo Suzuki**[§] • **Masaru Tomita**

[1] Institute for Advanced Biosciences, Keio University, Fujisawa 252-8520, Japan
[§] First two authors are equal contributors
Corresponding author: * gaou@sfc.keio.ac.jp

- Sequence conservation
- Replication strand skew
- Base composition
- Amino acid usage
- Codon usage

W700–W705 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 3 May 2010*
doi:10.1093/nar/gkq315

G-language genome analysis environment with REST and SOAP web service interfaces

Kazuharu Arakawa*, Nobuhiro Kido, Kazuki Oshita and Masaru Tomita

Institute for Advanced Biosciences, Keio University, Fujisawa 252-8520, Japan

- Base URL
  - http://rest.g-language.org/

methods

- data access methods
  - http://rest.g-language.org/method_list/gb
- analysis methods
  - http://rest.g-language.org/method_list
- documentation for 'load' method
  - http://rest.g-language.org/help/load

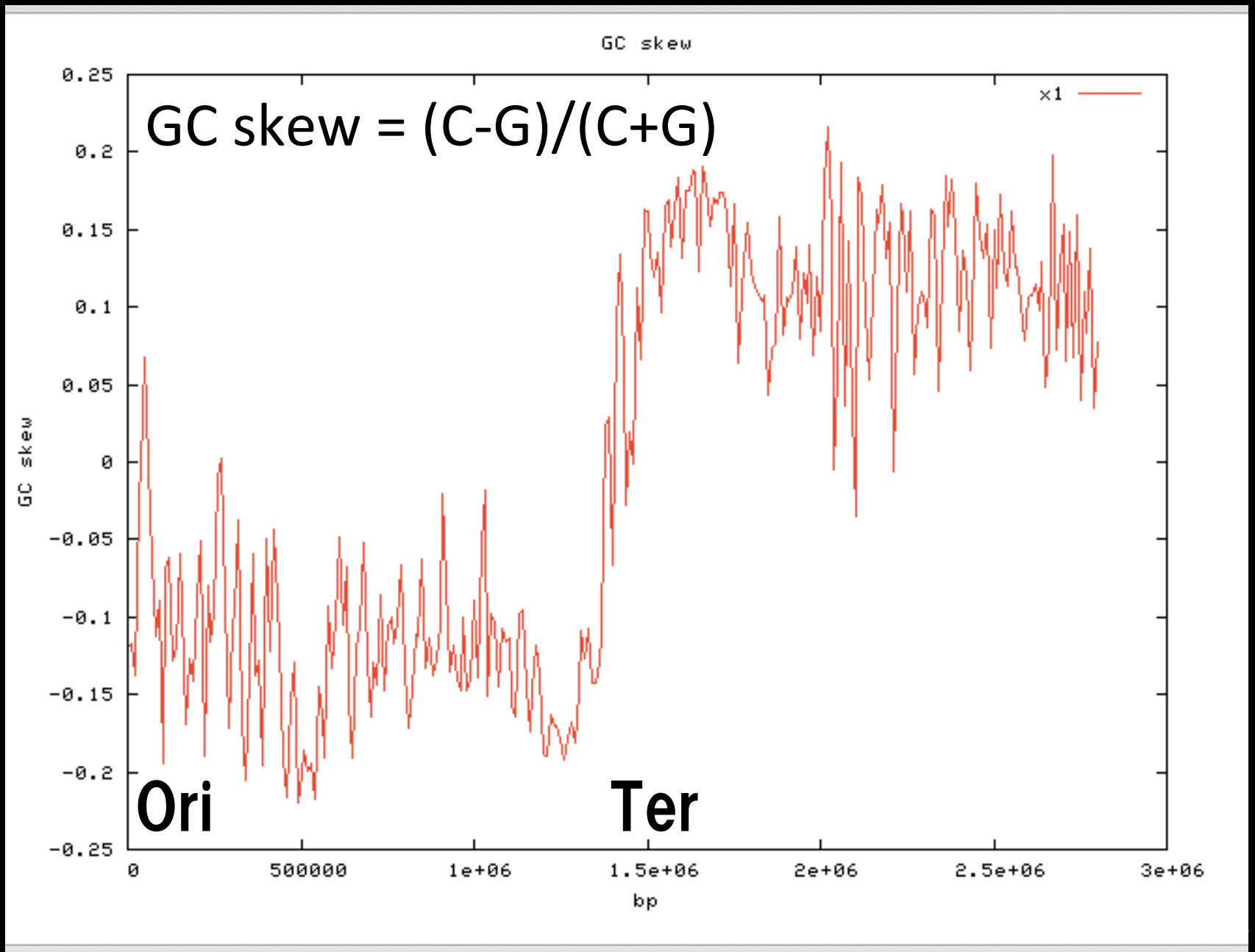
data

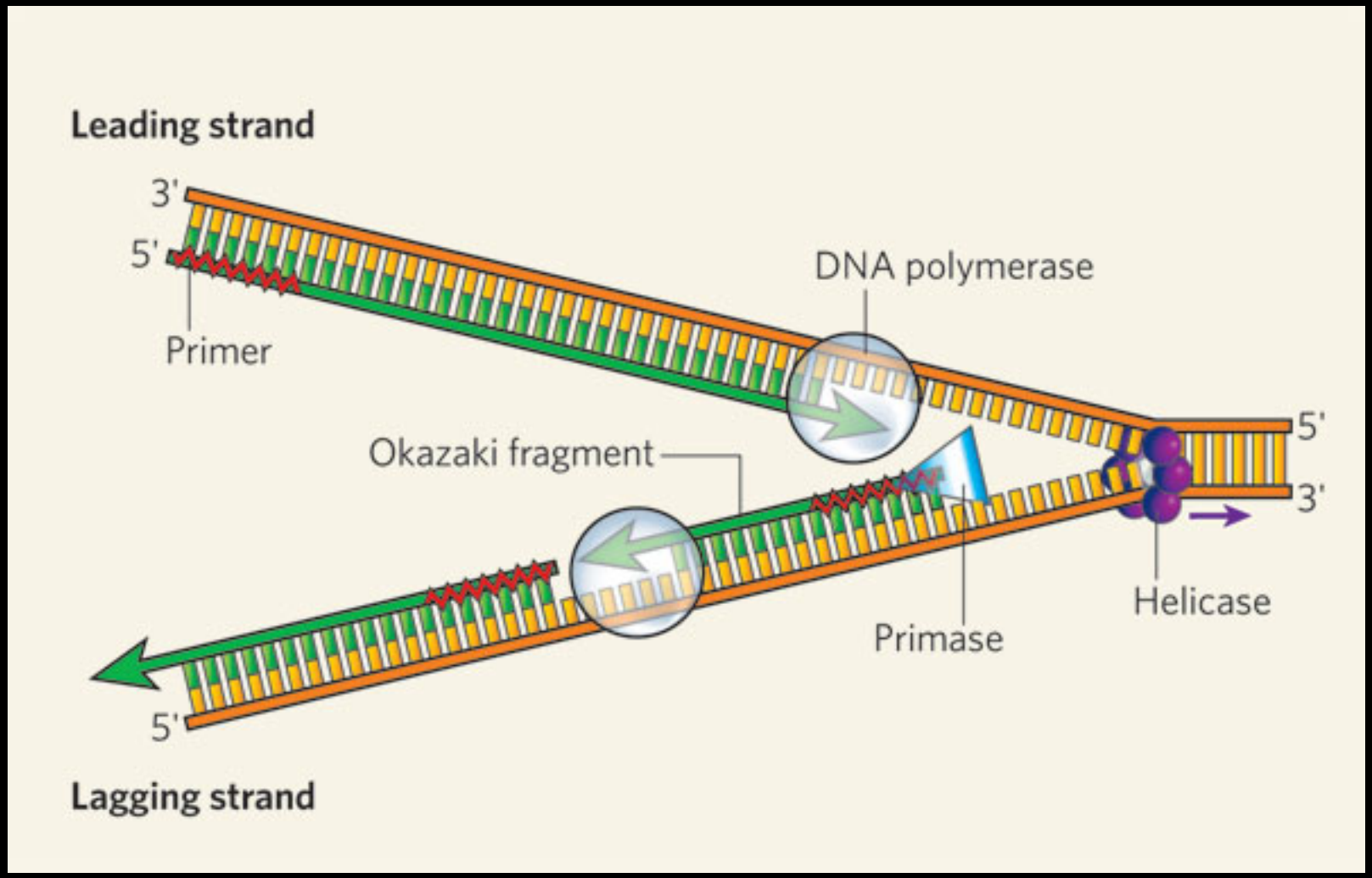
- list of available genomes
  - http://rest.g-language.org/organism_list/
- File upload
  - http://rest.g-language.org/upload/
- For more details, see Tutorial
  - http://www.g-language.org/wiki/restgenomeanalysisenglish

Genome analysis

- Replication strand skew
- Dinucleotide composition
- Codon usage

http://rest.g-language.org/NC_002745/gcskew

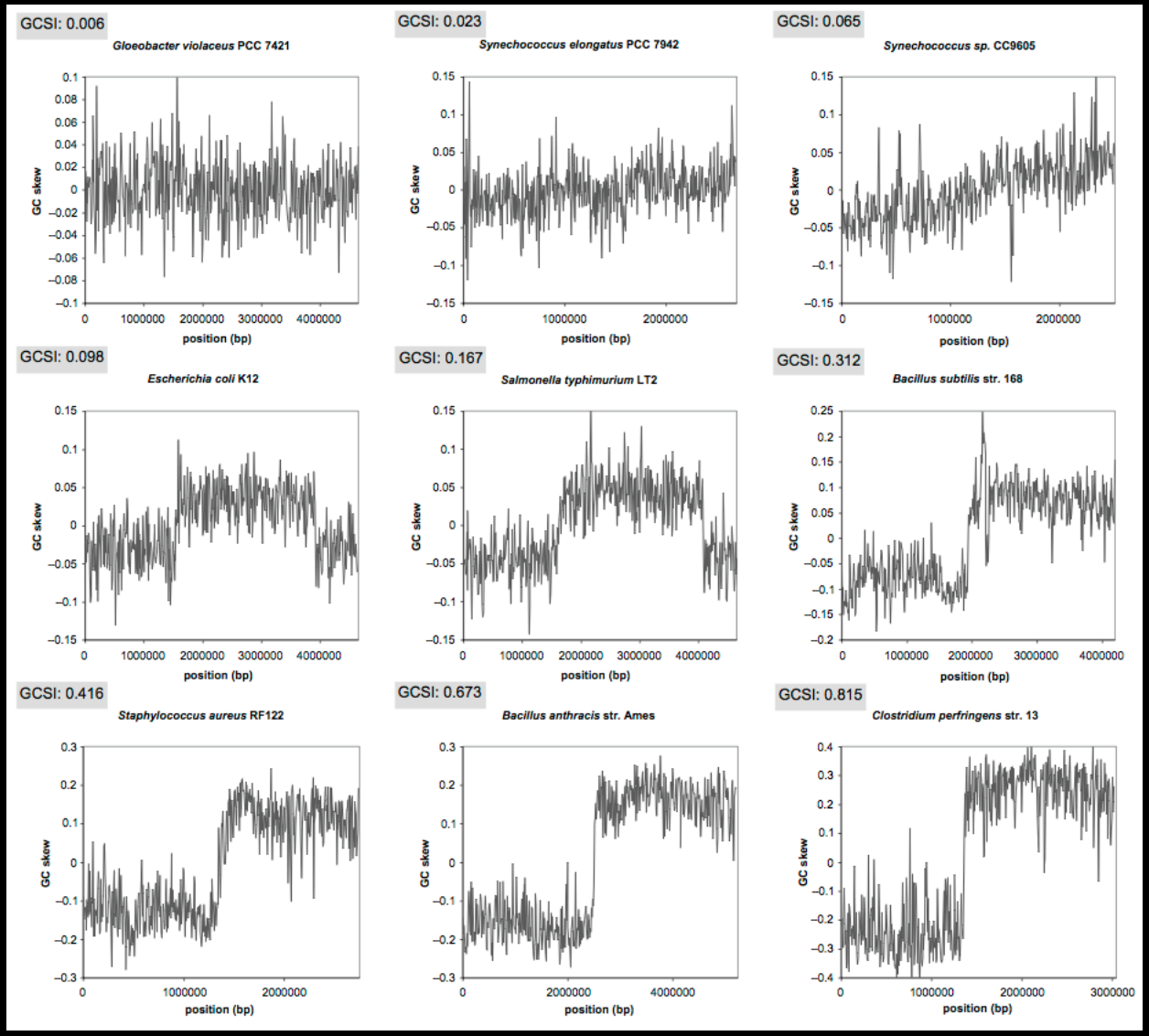




http://www.nature.com/scitable/content/dna-replication-of-the-leading-and-lagging-14668888

http://rest.g-language.org/NC_002745/gcsi

GC skew index (GCSI)



Arakawa K, Tomita M. Evol Bioinform Online. 2007 Sep 6;3:159-68.

Genome analysis

- Replication strand skew
- Dinucleotide composition
- Codon usage

Dinucleotide relative abundance

Genome signature

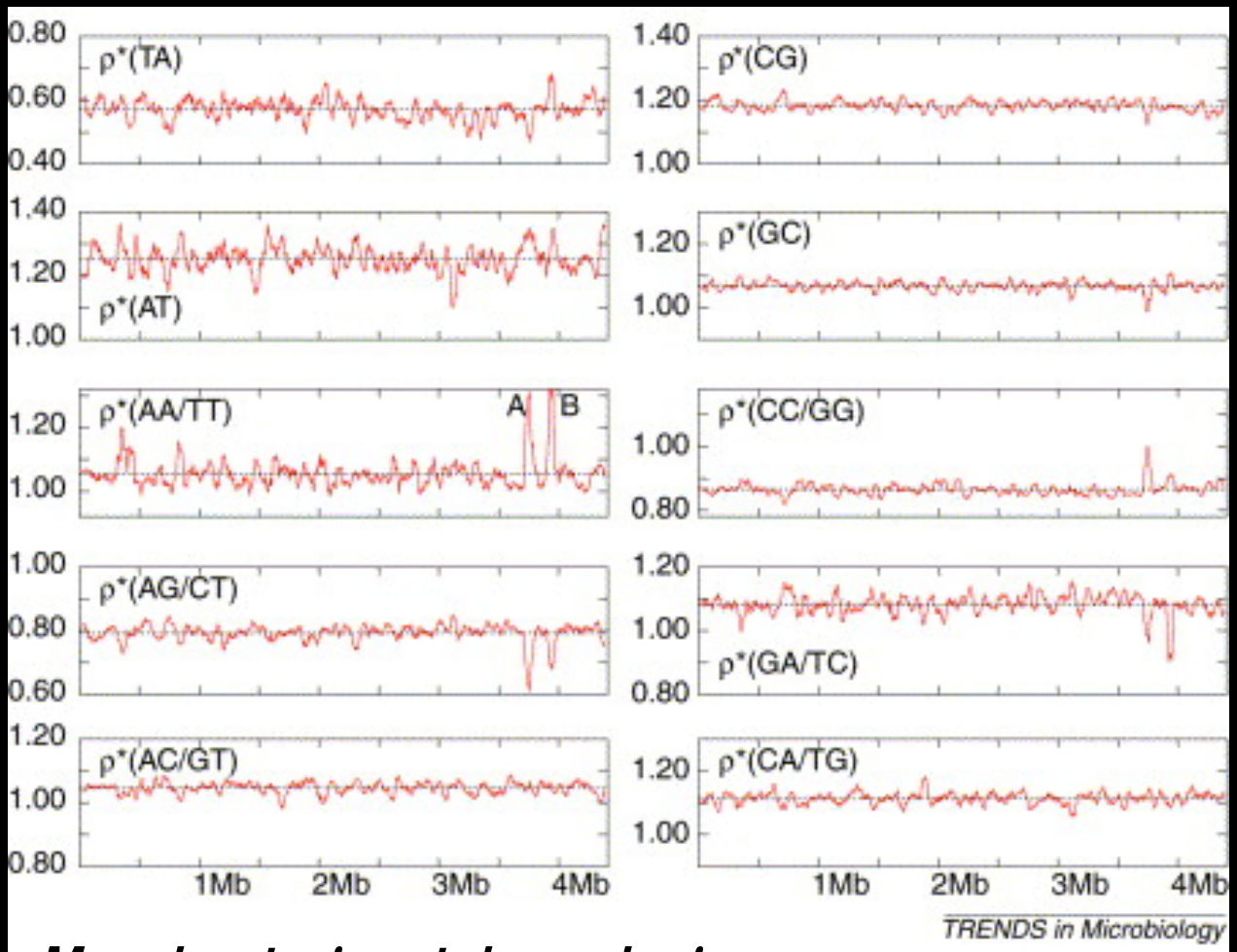
http://rest.g-language.org/NC_000956/signature

genome (available DNA)	CG	GC	TA	AT	CC GG	TT AA	TG CA	AG CT	AC GT	GA TC	G+C
Escherichia coli (4.6Mb)*	1.16	1.28	0.75	1.10	0.91	1.21	1.12	0.82	0.88	0.92	51%
Haemophilus influenzae (1.8Mb)*	1.09	1.43	0.75	0.95	1.01	1.25	1.12	0.82	0.85	0.87	38%
Neisseria gonorrhoeae (877kb)	1.32	1.26	0.63	1.05	0.99	1.50	0.99	0.67	0.83	0.89	53%
Neisseria meningitidis (2.2Mb)	1.31	1.27	0.64	1.05	0.96	1.44	1.01	0.70	0.84	0.91	52%
Rhodobacter capsulatus (1.4Mb)	1.19	1.19	0.33	1.61	0.88	1.30	1.03	0.84	0.71	1.16	67%
Rickettsia prowazekii (1.1Mb)*	0.77	1.53	0.98	0.98	1.03	1.05	1.02	1.06	0.86	0.91	29%
Helicobacter pylori (1.7Mb)*	0.93	1.56	0.73	0.86	1.17	1.37	0.97	0.97	0.67	0.87	39%
Campylobacter jejuni (1.6Mb)*	0.62	1.75	0.77	0.83	1.11	1.25	1.03	1.09	0.71	0.92	31%
Bacillus subtilis (4.2Mb)*	1.04	1.27	0.65	1.02	0.97	1.24	1.08	0.91	0.75	1.06	44%
Streptococcus pyogenes (985kb)	0.71	1.19	0.76	0.89	1.04	1.17	1.12	1.04	0.86	0.99	39%
Clostridium acetobutylicum (4.0Mb)	0.45	1.28	0.93	0.95	1.22	1.08	1.02	1.12	0.81	0.97	31%
Streptomyces coelicolor (2.4Mb)	1.14	0.97	0.51	0.93	0.88	0.82	1.00	0.95	1.14	1.25	72%
Mycobacterium leprae (1.7Mb)	1.13	1.07	0.75	1.10	0.88	1.04	1.14	0.86	1.05	1.02	58%
Mycobacterium tuberculosis (4.4Mb)*	1.18	1.07	0.58	1.24	0.86	1.05	1.11	0.80	1.05	1.08	65%
Mycoplasma genitalium (580kb)*	0.39	1.19	0.75	0.77	1.13	1.23	1.16	1.06	0.96	0.89	32%
Mycoplasma pneumoniae (816kb)*	0.82	1.14	0.77	0.71	1.12	1.30	1.08	0.96	1.02	0.81	40%
Synechocystis sp. (3.6Mb)*	0.75	1.02	0.75	1.00	1.36	1.32	1.05	0.85	0.79	0.86	48%
Deinococcus radiodurans (3.0Mb)	1.07	1.16	0.49	0.89	0.87	1.24	1.12	1.00	0.93	1.01	67%
Treponema pallidum (1.1Mb)*	1.08	1.22	0.74	0.93	0.86	1.18	1.13	0.94	0.96	0.95	53%
Borrelia burgdorferi (911kb)*	0.48	1.47	0.77	0.88	1.29	1.22	1.02	1.07	0.69	1.01	29%
Chlamydia trachomatis (1.0Mb)*	0.79	1.12	0.77	0.89	1.01	1.16	0.96	1.18	0.76	1.15	41%
Aquifex aeolicus (1.6Mb)*	0.87	0.75	0.82	0.66	1.24	1.29	0.74	1.18	0.89	1.12	43%

Campbell A et al. PNAS 1999;96:9184-9189

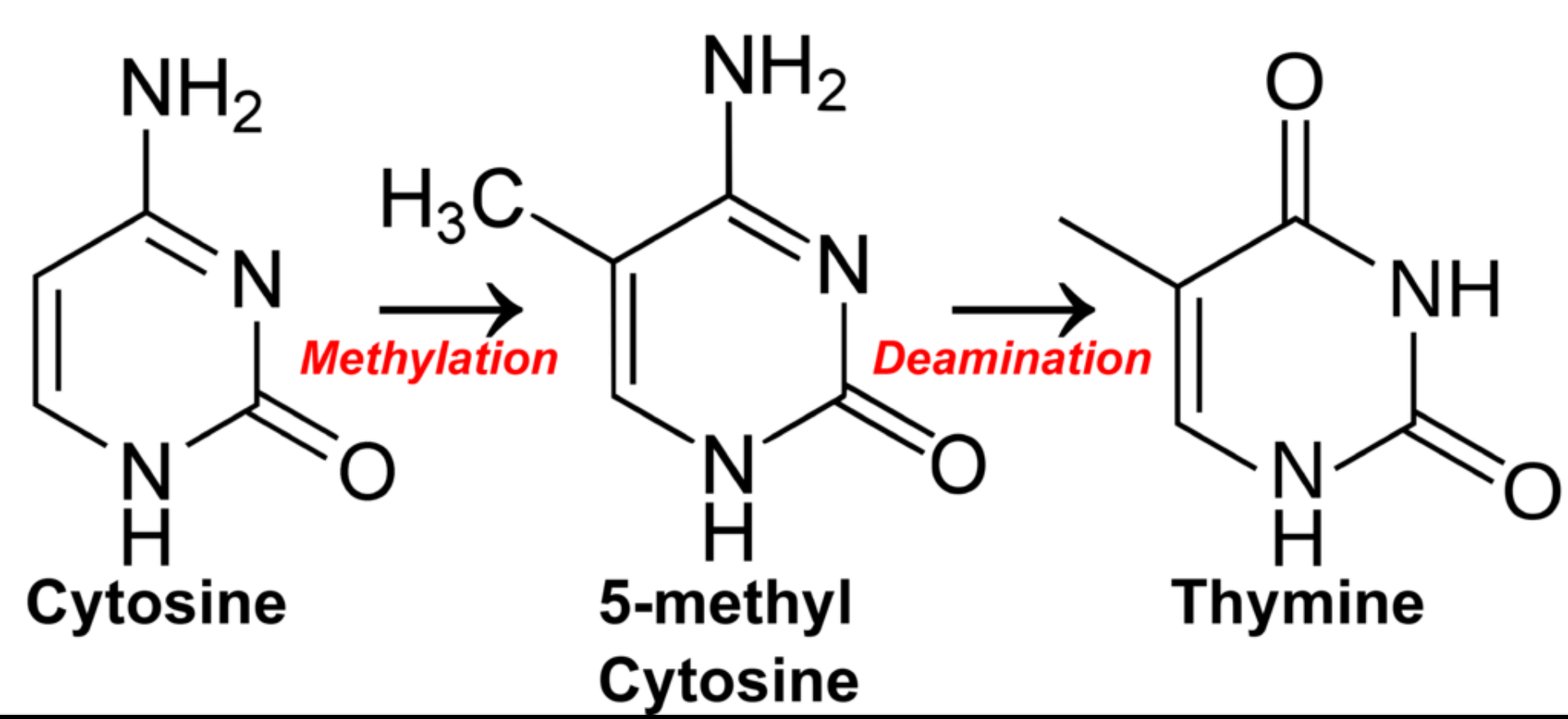
Dinucleotide relative abundance

Genome signature



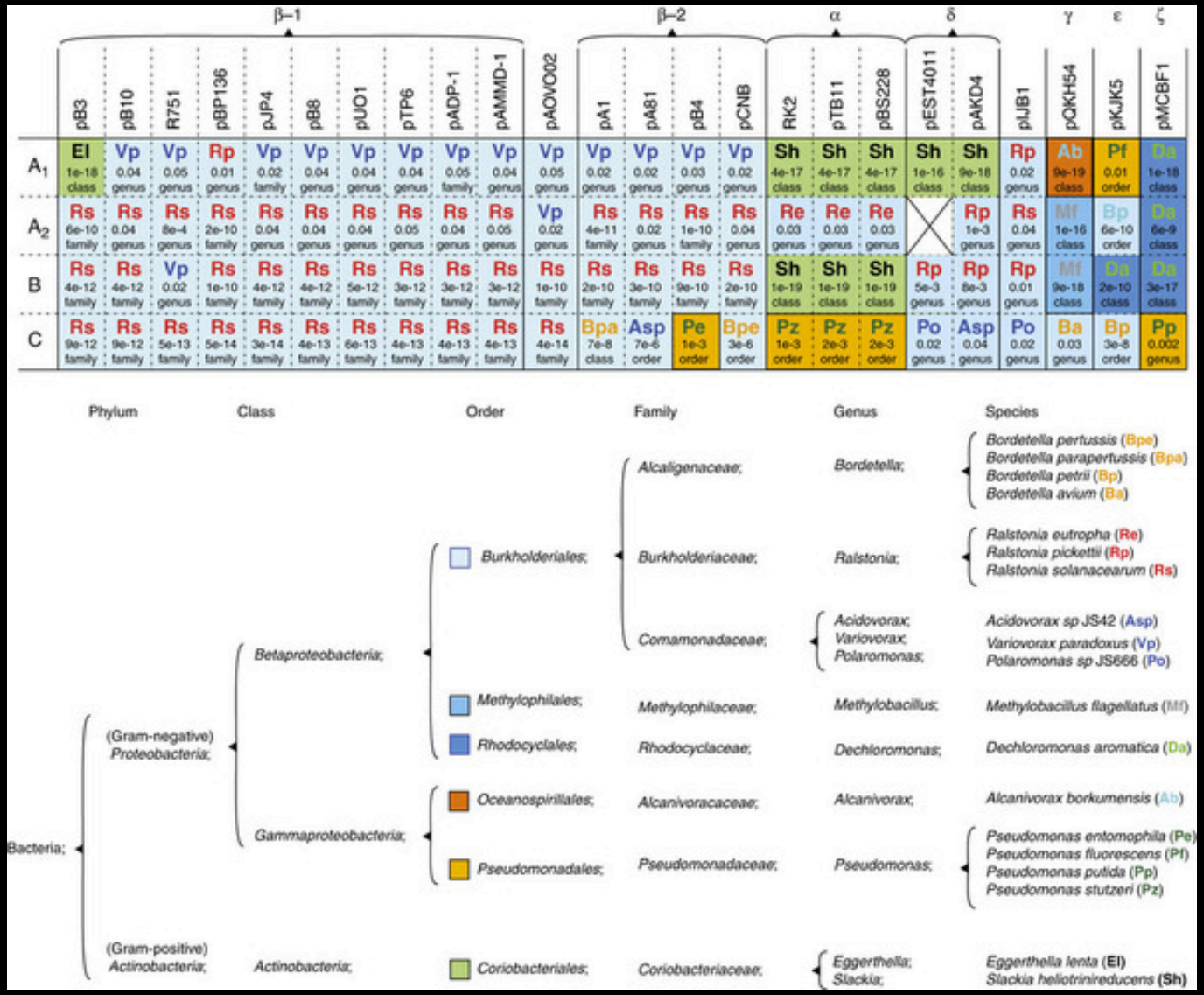
Mycobacterium tuberculosis **genome**

Karlin S. Trends Microbiol. 2001;9(7):335-43



http://en.wikipedia.org/wiki/CpG_site

Taxonomic classification: Predicting plasmid hosts



	β–1											β–2				α			δ			γ	ε	ζ
	pB3	pB10	R751	pBP136	pJP4	pB8	pUO1	pTP6	pADP-1	pAMMD-1	pAOV002	pA1	pA81	pB4	pCNB	RK2	pTB11	pBS228	pEST4011	pAKD4	pJB1	pQKH54	pKJK5	pMCBF1
A_1	El 1e-18 class	Vp 0.04 genus	Vp 0.05 genus	Rp 0.01 genus	Vp 0.02 family	Vp 0.04 genus	Vp 0.04 genus	Vp 0.04 genus	Vp 0.05 family	Vp 0.04 genus	Vp 0.05 genus	Vp 0.02 genus	Vp 0.02 genus	Vp 0.03 genus	Vp 0.02 genus	Sh 4e-17 class	Sh 4e-17 class	Sh 4e-17 class	Sh 1e-16 class	Sh 9e-18 class	Rp 0.02 genus	Ab 9e-19 class	Pf 0.01 order	Da 1e-18 class
A_2	Rs 6e-10 family	Rs 0.04 genus	Rs 8e-4 genus	Rs 2e-10 family	Rs 0.04 genus	Rs 0.04 genus	Rs 0.04 genus	Rs 0.05 genus	Rs 0.04 genus	Rs 0.05 genus	Vp 0.02 genus	Rs 4e-11 family	Rs 0.02 genus	Rs 1e-10 family	Rs 0.04 genus	Re 0.03 genus	Re 0.03 genus	Re 0.03 genus		Rp 1e-3 genus	Rs 0.04 genus	Mf 1e-16 class	Bp 6e-10 order	Da 6e-9 class
B	Rs 4e-12 family	Rs 4e-12 family	Vp 0.02 genus	Rs 1e-10 family	Rs 4e-12 family	Rs 4e-12 family	Rs 5e-12 family	Rs 3e-12 family	Rs 3e-12 family	Rs 3e-12 family	Rs 1e-10 family	Rs 2e-10 family	Rs 3e-10 family	Rs 9e-10 family	Rs 2e-10 family	Sh 1e-19 class	Sh 1e-19 class	Sh 1e-19 class	Rp 5e-3 genus	Rp 8e-3 genus	Rp 0.01 genus	Mf 9e-18 class	Da 2e-10 class	Da 3e-17 class
C	Rs 9e-12 family	Rs 9e-12 family	Rs 5e-13 family	Rs 5e-14 family	Rs 3e-14 family	Rs 4e-13 family	Rs 6e-13 family	Rs 4e-13 family	Rs 4e-13 family	Rs 4e-13 family	Rs 4e-14 family	Bpa 7e-8 class	Asp 7e-6 order	Pe 1e-3 order	Bpe 3e-6 order	Pz 1e-3 order	Pz 2e-3 order	Pz 2e-3 order	Po 0.02 genus	Asp 0.04 genus	Po 0.02 genus	Ba 0.03 genus	Bp 3e-8 order	Pp 0.002 genus

Norberg P et al. Nat Commun. 2011;2:268.

Genome analysis

- Replication strand skew
- Dinucleotide composition
- Codon usage

- Synonymous codon usage is related to gene expression, replication, GC content, etc.

The Standard Genetic Code

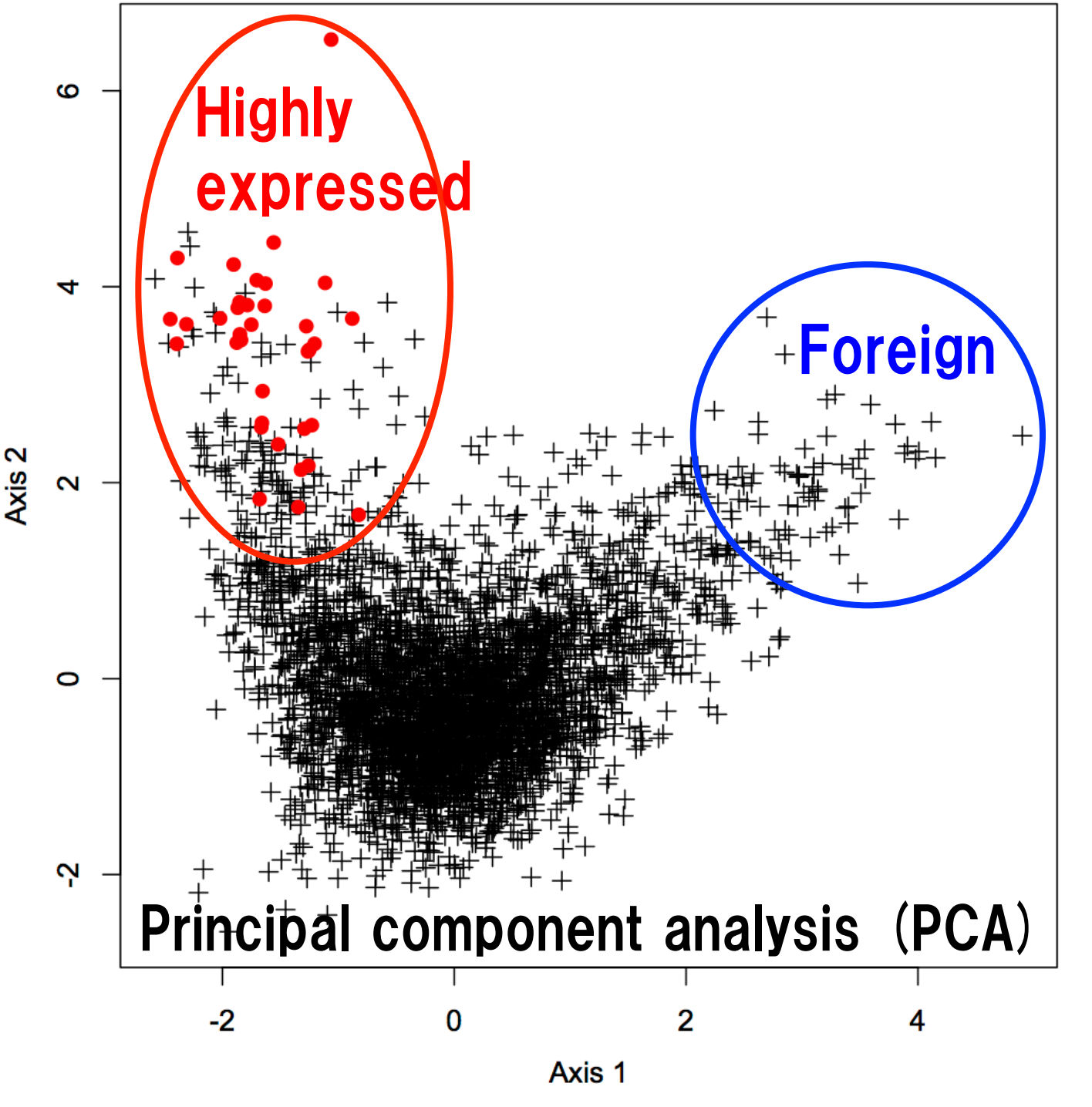
	T			C			A			G		
T	TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C
	TTC			TCC			TAC			TGC		
	TTA	Leu	L	TCA			TAA	STOP		TGA	STOP	
	TTG			TCG			TAG			TGG	Trp	W
C	CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R
	CTC			CCC			CAC			CGC		
	CTA			CCA			CAA	Gln	Q	CGA		
	CTG			CCG			CAG			CGG		
A	ATT	Ile	I	ACT	Thr	T	AAT	Asn	N	AGT	Ser	S
	ATC			ACC			AAC			AGC		
	ATA			ACA			AAA	Lys	K	AGA	Arg	R
	ATG	Met	M	ACG			AAG			AGG		
G	GTT	Val	V	GCT	Ala	A	GAT	Asp	D	GGT	Gly	G
	GTC			GCC			GAC			GGC		
	GTA			GCA			GAA	Glu	E	GGA		
	GTG			GCG			GAG			GGG		

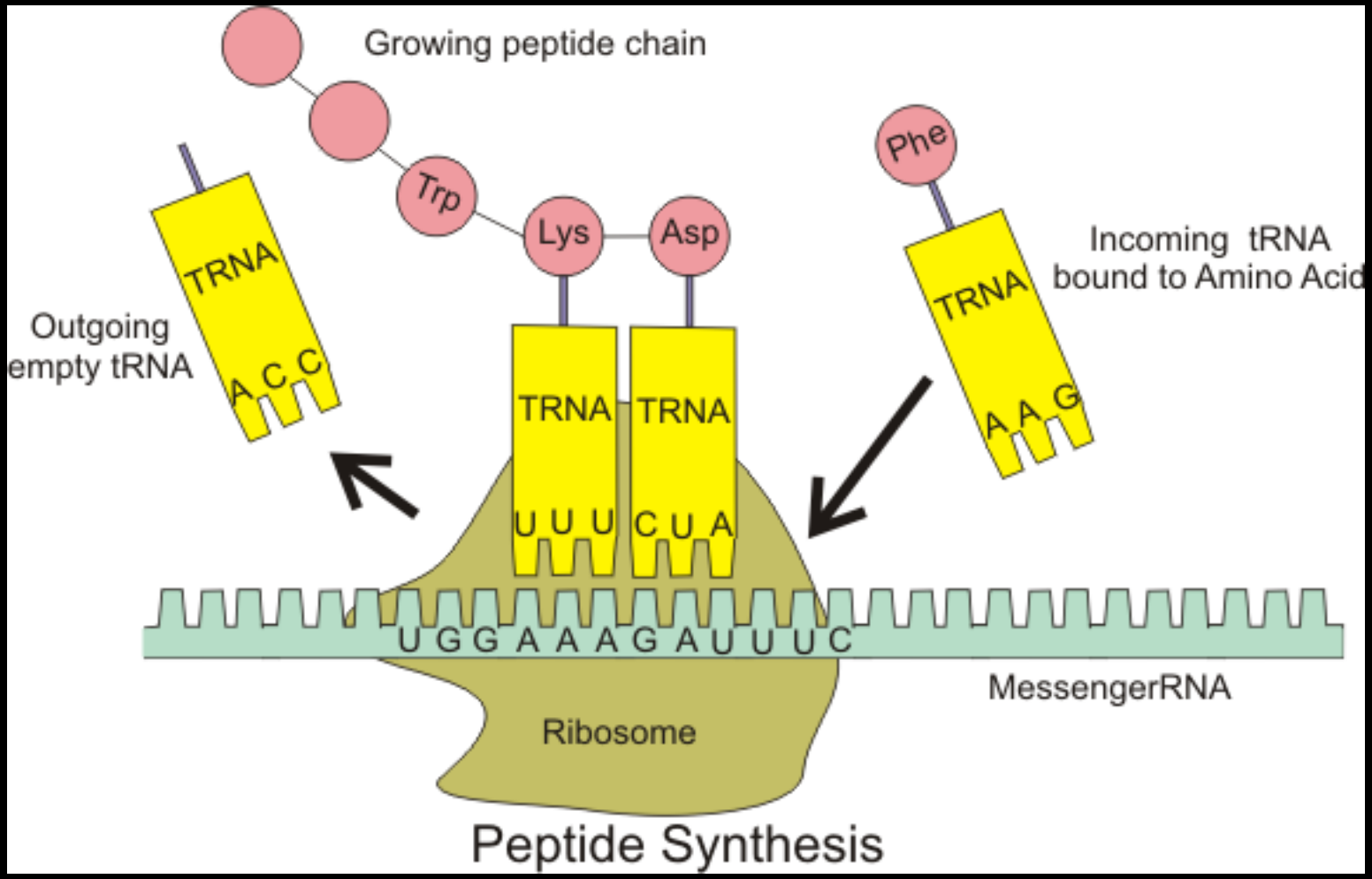
http://scienceblogs.com/digitalbio/2009/04/30/blasting-through-the-flu-activ/

http://rest.g-language.org/NC_000956/codon_usage

first	second												third
	T			C			A			G			
T	TTT	F	0.857	TCT	S	0.251	TAT	Y	0.725	TGT	C	0.678	T
	TTC	F	0.143	TCC	S	0.052	TAC	Y	0.275	TGC	C	0.322	C
	TTA	L	0.425	TCA	S	0.216	TAA	/	0.639	TGA	/	0.111	A
	TTG	L	0.111	TCG	S	0.044	TAG	/	0.250	TGG	W	1.000	G
C	CTT	L	0.267	CCT	P	0.355	CAT	H	0.756	CGT	R	0.111	T
	CTC	L	0.028	CCC	P	0.143	CAC	H	0.244	CGC	R	0.040	C
	CTA	L	0.138	CCA	P	0.430	CAA	Q	0.865	CGA	R	0.083	A
	CTG	L	0.032	CCG	P	0.072	CAG	Q	0.135	CGG	R	0.018	G
A	ATT	I	0.524	ACT	T	0.428	AAT	N	0.791	AGT	S	0.304	T
	ATC	I	0.084	ACC	T	0.121	AAC	N	0.209	AGC	S	0.133	C
	ATA	I	0.392	ACA	T	0.401	AAA	K	0.825	AGA	R	0.609	A
	ATG	M	0.911	ACG	T	0.050	AAG	K	0.175	AGG	R	0.139	G
G	GTT	V	0.460	GCT	A	0.399	GAT	D	0.761	GGT	G	0.288	T
	GTC	V	0.059	GCC	A	0.094	GAC	D	0.239	GGC	G	0.167	C
	GTA	V	0.373	GCA	A	0.428	GAA	E	0.775	GGA	G	0.383	A
	GTG	V	0.108	GCG	A	0.078	GAG	E	0.225	GGG	G	0.162	G

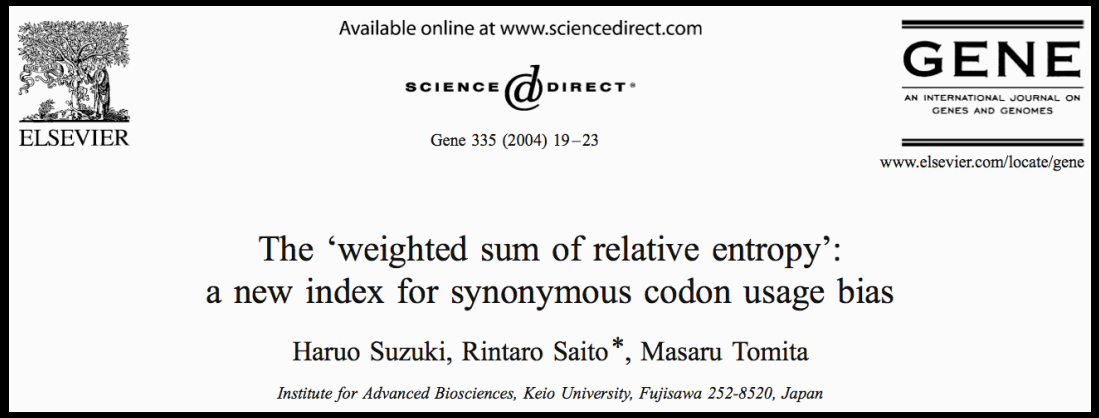
yellow minus charge red plus charge blue noncharge green nonpolar


Highly expressed
Foreign
Principal component analysis (PCA)
Axis 2
Axis 1
6
4
2
0
-2
-2
0
2
4



http://upload.wikimedia.org/wikipedia/commons/0/0f/Peptide_syn.png

Proposed methods



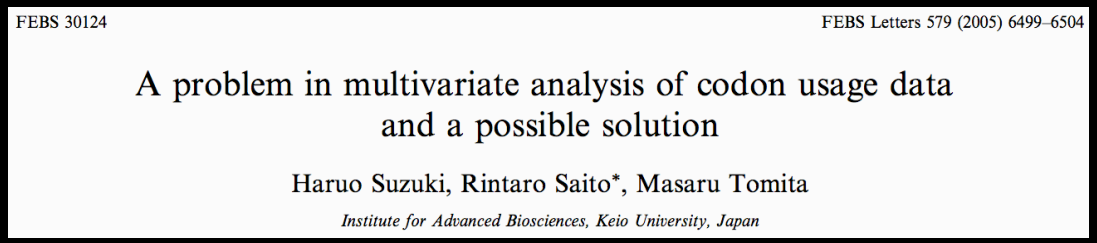
Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 335 (2004) 19–23
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
ELSEVIER
www.elsevier.com/locate/gene

The 'weighted sum of relative entropy': a new index for synonymous codon usage bias

Haruo Suzuki, Rintaro Saito*, Masaru Tomita

Institute for Advanced Biosciences, Keio University, Fujisawa 252-8520, Japan

http://rest.g-language.org/NC_000956/Ew



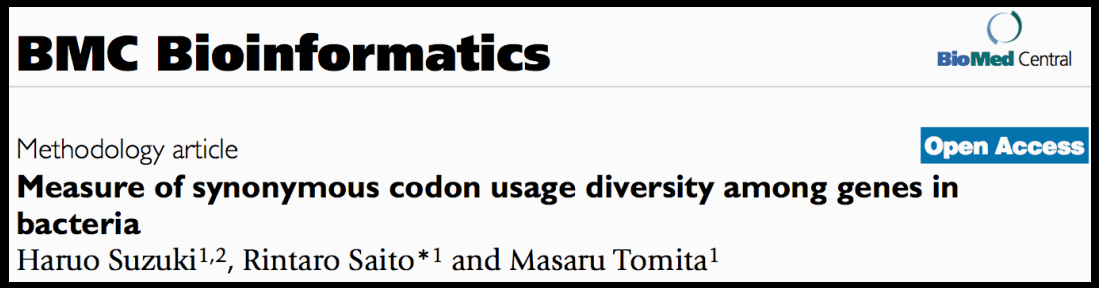
FEBS 30124
FEBS Letters 579 (2005) 6499–6504

A problem in multivariate analysis of codon usage data and a possible solution

Haruo Suzuki, Rintaro Saito*, Masaru Tomita

Institute for Advanced Biosciences, Keio University, Japan

http://rest.g-language.org/NC_000956/codon_mva



BMC Bioinformatics
BioMed Central
Methodology article
Open Access

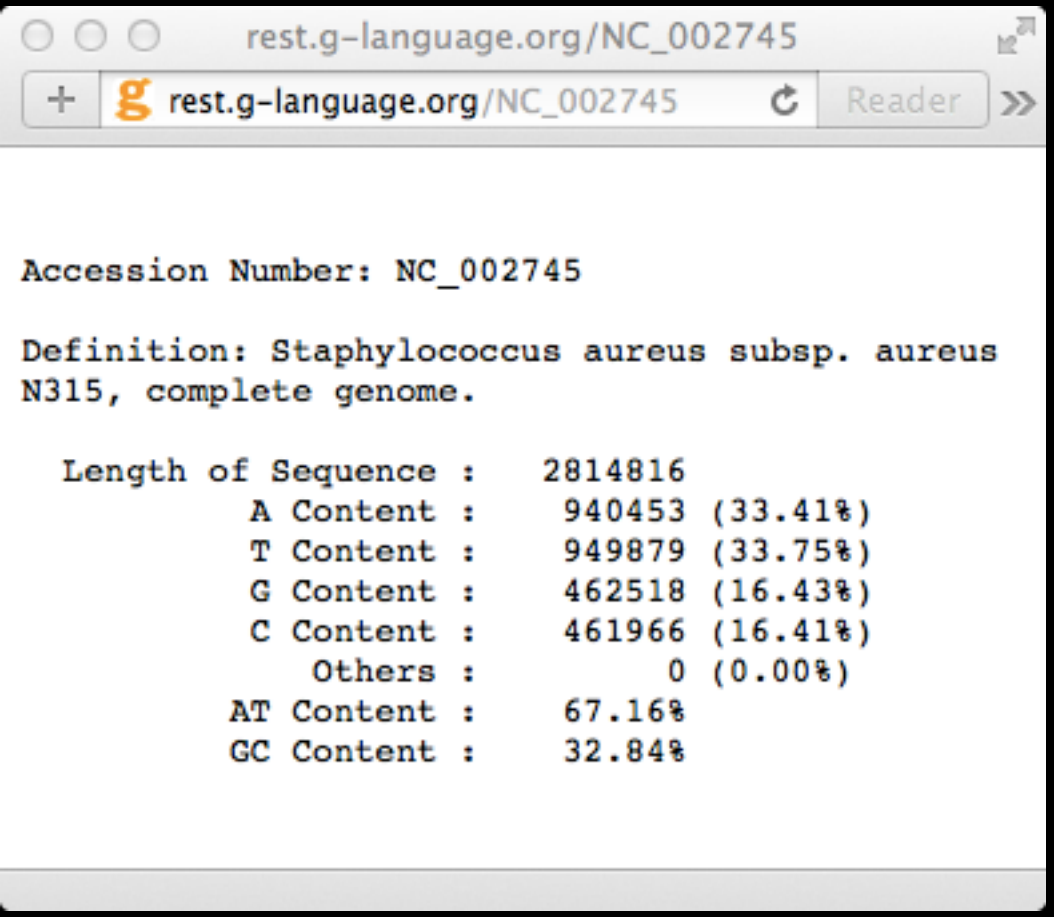
Measure of synonymous codon usage diversity among genes in bacteria

Haruo Suzuki[1,2], Rintaro Saito*[1] and Masaru Tomita[1]

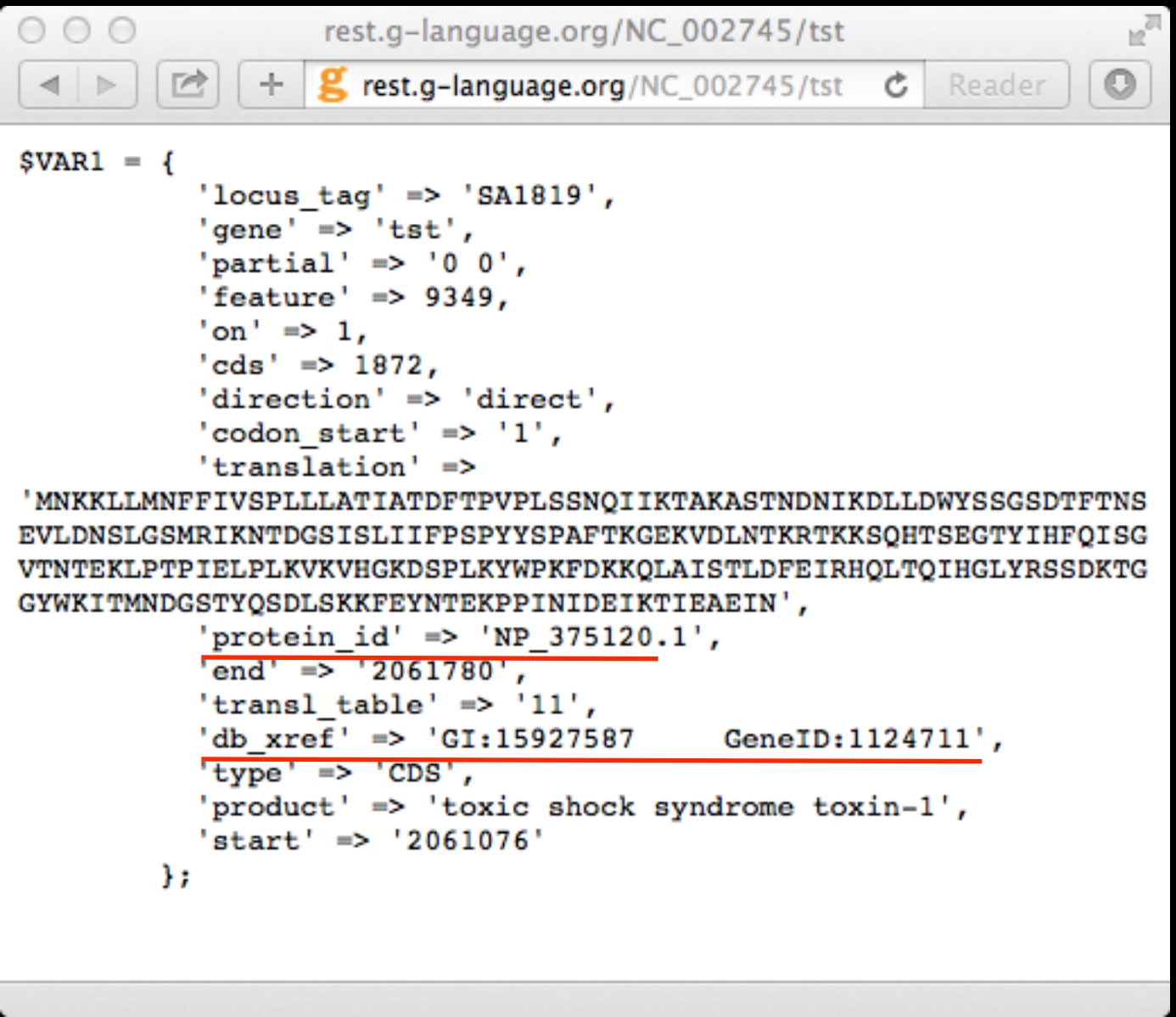
http://rest.g-language.org/NC_000956/Dmean

Demo

http://rest.g-language.org/NC_002745
show information for *Staphylococcus aureus* N315 genome (accession: NC_002745)



http://rest.g-language.org/NC_002745/tst
show information for *tst* gene



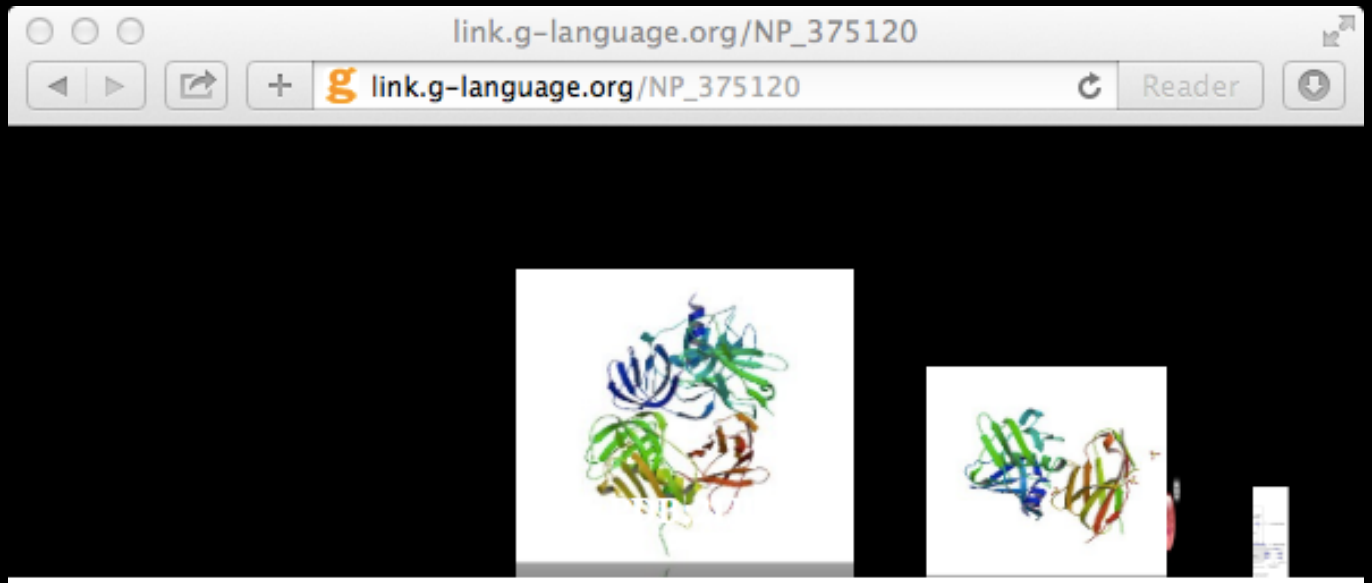
G-Links

collects related information to a given gene.

Base URL

- http://link.g-language.org/

http://link.g-language.org/NP_375120
http://link.g-language.org/GI:15927587
http://link.g-language.org/GeneID:1124711



DataBase	ID	URL or Descriptions
# Gene3D	3.10.20.120	-; 1.
# GO_component	GO:0005576	extracellular region; IEA:InterPro.
# GO_process	GO:0009405	pathogenesis; IEA:InterPro.
# GOslim_component	GO:0005576	extracellular region
# GOslim_process	GO:0008150	biological_process
# InterPro	IPR006123	Toxin_b-grasp_Staph/Strep
# InterPro	IPR006125	Staph_toxin

Sample genome data

ACCESSION	Genome	(preset name)
NC_000913	Escherichia coli K12 MG1655	(ecoli)
NC_000964	Bacillus subtilis	(bsub)
NC_000908	Mycoplasma genitalium	(mgen)
NC_005070	Synechococcus sp.	(cyano)
NC_003413	Pyrococcus furiosus	(pyro)
NC_001318	Borrelia burgdorferi B31	(bbur)
NC_002483	Plasmid F	(plasmidf)
NC_001416	Enterobacteria phage lambda	(lambda)

Examples

http://rest.g-language.org/ecoli

http://rest.g-language.org/ecoli/recA

http://rest.g-language.org/ecoli/*/product

http://rest.g-language.org/ecoli/base_entropy

http://rest.g-language.org/ecoli/gcskew

http://rest.g-language.org/ecoli/gcwin

http://rest.g-language.org/ecoli/codon_usage

http://rest.g-language.org/ecoli/cai/tag=gene