Orthologous and paralogous FFRPs in *E. coli* and related proteobacteria

By Katsushi YOKOYAMA*), **) and Masashi SUZUKI*), †)

(Communicated by Masanori OTSUKA, M. J. A.)

Abstract: Feast/famine regulatory proteins (FFRP) comprise a diverse family of transcription factors. Orthologues of types of eubacterial FFRPs, Lrp, AsnC, YbaO and TinR, were identified. Organisms having these FFRPs were found limited into smaller groups in succession: Lrp in the β and γ subclasses of *Proteobacteria*, AsnC in the facultatively anaerobic order in the γ subclass, YbaO in the same order in its *Vibrionaceae* and *Enterobacteriaceae* families only, and TinR in the species *Salmonella enterica* in the *Enterobacteriaceae* family. Yet in a distance map (i.e. an unrooted phylogenetic tree), e.g. TinR did not branch from the YbaO cluster, but the four FFRP groups remained outside to each other. These facts can be explained by assuming duplication of the ancestor gene of *ybaO* and *tinR* inside a common ancestor of *Vibrionaceae* and *Enterobacteriaceae*. One of the two genes was modified to *ybaO*, and the other was unused for a number of years until it adapted the new function of *tinR* inside *S. enterica* after diversification of *Vibrionaceae* and *Enterobacteriaceae*. Using various outgroups, the common root of the four FFRP types was localized to the connection between AsnC and the other three types, so that the second differentiation took place between Lrp and the common ancestor of YbaO and TinR.

Key words: Asparagine synthase C gene product (AsnC); evolution; leucine responsive regulatory protein (Lrp); neutral evolution; protein evolution; protein family; transcription regulation.

Introduction: the nature of a distance map. Since the classic work of Zuckerkandl and Pauling,[1] the amino acid sequences of sets of proteins orthologous to each other, e.g. myoglobins from various mammals, have been extensively used in order to clarify the phylogenetic relation of organisms. Differences between pairs of amino acid sequences are expressed as distances, and a type of a topological map is made, so that it best satisfies these distances: here referred to as a distance map or an unrooted phylogenetic tree. Use of orthologous sequences can be rationalized by assuming an accumulation of random mutations, which has not affected the original function of the protein, thereby recording the history of divergence: neutral evolution.[2]

Ideally, amino acid positions in a protein can be classified into three types: the first type of positions where amino acids are unchangeable, since they are essential for the function of the protein; the second type where amino acids are selected for adapting to the natural history of the organism e.g. its surviving temperature, high or low, or its body size, large or small: and the third type where different amino acid residues are tolerated.

Differences at positions in the third type are useful for identifying the phylogeny of the source organisms. For this purpose, conserved positions in the first type are not useful, but they can be used for identification of proteins orthologous to each other. Positions of hemoglobins, which are important for producing a cooperativity in binding to oxygen molecules, can be used to separate hemoglobins from myoglobins. Second type positions are important for understanding designs of proteins, but can mislead phylogenetic identification. Hemoglobins in large and small mammals have different oxygen-binding constants, Bohr effects *etc*.[3] If a single strategy only is available to modify the protein upon the change in the body size, amino acid residues at some positions will be kept the same in hemoglobins of an elephant and a dinosaur, which are not closely related: i.e. convergence at the molecular level. Other problems such as double mutations at the same position can also compli-

*) National Institute of Advanced Industrial Science and Technology (AIST), AIST Tsukuba Center 6-10, Higashi 1-1-1, Tsukuba, Ibaraki 305-8566, Japan.

**) Japan Science and Technology Agency (JST), Core Research for Evolutionary Science and Technology (CREST), Honmachi 4-1-18, Kawaguchi Center Building, Kawaguchi, Saitama 332-0012, Japan.

†) Correspondence to: M. Suzuki.

cate the situation.

Since it is unknown what position belongs to which type, usually bootstrapping is carried out. Bootstrapping is a process of making a series of distance maps, imposing different weights on amino acid positions. At each node separating pairs of organisms, the number of times is represented, when the same result is obtained. Only when the number of singly mutating residues in the third type is large, will a high bootstrapping value be deduced.

Information obtainable from a distance map is not necessarily limited to the phylogeny of organisms. As will be described in this paper, by mapping a combination of protein sequences orthologous or paralogous to each other the process of gene duplication and differentiation of proteins can be understood. Often such a process is called "protein evolution".[4] However, we try to avoid use of this term, since evolution is a process of selecting an organism as a system, and the pressure does not directly apply to individual components.

Proteins studied in this paper. Orthologous and paralogous proteins which we study in this paper are FFRPs (feast/famine regulatory proteins).[5)-22)] These transcription factors regulate metabolic pathways in archaea and eubacteria. In order to understand the original form of transcription regulation in the common ancestor of all the extant organisms, and the history of its modification, which has enabled evolution of the ancestor to extant organisms, gene duplication and functional differentiation of FFRPs need to be understood.

In *E. coli* three FFRPs are present. Lrp (leucine responsive regulatory protein) activates or represses transcription of a number of genes of *E. coli*, in many cases depending on extra-cellular leucine.[23),24)] When sensing rich nutrition as a high concentration of leucine in the environment, *E. coli* changes its metabolism, terminating autotrophic pathways and activating absorption of nutrients *via* the cell membrane, thereby shifting to a more heterotrophic mode and altering its infectivity by regulating formation of pili.

The protein AsnC (asparagine synthase C gene product) is a paralogue of Lrp, as hemoglobin is of myoglobin. Depending on the concentration of asparagine, AsnC down-regulates autotrophic biosynthesis of this amino acid.[25)] Positioning of the replication origin (*oriC*) of *E. coli* inside the *asn* operon[26)] hints at involvement of AsnC in cell replication. A third *E. coli* FFRP, YbaO, was identified using the genomic sequence, and its function remains unknown. A related enterobacterium, *Salmonella enterica*, has a fourth FFRP, TinR.[27)]

Requirements for orthologues of *E. coli* and *S. enterica* FFRPs. Orthologues of the four types of FFRPs were identified (Table I) using the Blast program[28)] and the NCBI database (http://www.ncbi.nlm.nih.gov/), so that they fulfilled the following requirements.

Firstly, these orthologues ought to fold into 3D structures essentially the same as seen in crystal structures of other FFRPs,[5),9),29)] by combining a particular set of secondary structural elements. Thus, amino acid sequences were carefully analyzed, by confirming e.g. the 3.6 periodicity made by hydrophobic residues inside regions expected to form α helices.[7),8),11)]

Secondly, an orthologue, e.g. of Lrp, ought to have higher homology to the original FFRP than to the three other paralogues, AsnC, YbaO, and TinR. Principally, a single protein should be assigned for each organism as an orthologue of a protein, and this principle was confirmed in this study.

Thirdly, an orthologue ought to have amino acid residues essentially the same as those in the *E. coli* or *S. enterica* FFRP inside its α helix 3 (Fig. 1). Different types of FFRPs will have different DNA-binding specificities, and their α helices 3 are used for discrimination between DNA promoters.[10)] In short, the α helix contains first type positions described in *Introduction*.

Lastly and importantly, in a distance map orthologous proteins ought to be related by a topology consistent with the known phylogeny of the source organisms (Figs. 2-4). In other words, in this study protein orthologues were identified by reversing the ordinary direction of "orthologues to phylogeny".

Orthologues of the *E. coli* and *S. enterica* FFRPs identified. All of the orthologues identified (Table I) were proteins from eubacteria in the class *Proteobacteria* (Fig. 2). Distance maps made using orthologues of the four types were consistent with each other (Figs. 2-4), and also with the standard phylogenetic tree in most respects.

The class *Proteobacteria* is divided into subclasses α-ε. Orthologues of Lrp were found in organisms in the β and γ subclasses (shown in blue in Fig. 2). Multiple types of FFRPs, having high homology to *E. coli* Lrp, which, however, were different in their amino acid sequences inside α helices 3 (Fig. 1), were found distributing mainly in the α subclass (shown in red in Fig. 2), thereby complementing the distribution of Lrp orthologues: α proteobacterial FFRPs (αPBFs). Using the amino acid sequences of αPBFs and Lrp orthologues, a distance

Table I. List of proteins studied in this work

Lrp No	Species	NCBI code	genome position
1	Chromobacterium violaceum ATCC12472	NP_901583	2078652
2	Neisseria meningitidis MC58 serogroup B	NP_274655	1717742
3	Neisseria meningitidis Z2491 serogroup A	NP_284612	1832690
4	Xanthomonas axonopodis pv.vesicatoria	AAP30679	–
5	Azotobater vinelandii	ZP_00092259	–
6	Pseudomonas aeruginosa PAO1	NP_253995	5977610
7	Pseudomonas putida KT2440	NP_747372	6021284
8	Pseudomonas fluorescens PfO-1	ZP_00264884	–
9	Pseudomonas syringae pv. tomato str .DC3000	NP_789959	120331
10	Pseudomonas syringae pv. syringae B728a	ZP_00124904	–
11	Burkholderia fungorum LB400	ZP_00281038	–
12	Burkholderia pseudomallei K96243	YP_109091	I-3008637
13	Burkholderia mallei ATCC23344	YP_102221	I-429896
14	Burkholderia cepacia R18194	ZP_00217274	–
15	Burkholderia cepacia R1808	ZP_00222570	–
16	Ralstonia solanacearum GMI1000	NP_519048	980078
17	Ralstonia eutropha JMP134	ZP_00168378	–
18	Ralstonia metallidurans CH34	ZP_00275405	–
19	Bordetella bronchiseptica RB50	NP_889586	3257127
19	Bordetella parapertussis 12822	NP_885266	3328916
19	Bordetella pertussis Tohama I	NP_881252	2792791
20	Acinetobacter sp. ADP1	YP_044913	115051
21	Idiomarina loihiensis L2TR	YP_155058	723670
22	Haemophilus somnus 2336	ZP_00132211	–
22	Haemophilus somnus 129PT	ZP_00122493	–
23	Pasteurella multocida subsp. multocida str. Pm70	NP_245191	285262
24	Manheimia succiniciproducens MBEL55E	YP_088647	1450734
25	Haemophilus influenzae 86-028NP	ZP_00321819	–
25	Haemophilus influenzae R2846	ZP_00155162	–
26	Haemophilus influenzae Rd KW20	NP_439738	1663262
27	Haemophilus influenzae R2866	ZP_00157113	–
28	Haemophilus ducreyi 35000HP	NP_873890	1213455
29	Actinobacillus pleuropneumoniae serovar 1 str. 4074	ZP_00135283	–
30	Shewanella oneidensis MR-1	NP_717900	2416772
31	Yersinia pestis biovar Medievalis str. 91001	NP_992584	1329782
31	Yersinia pestis CO92	NP_404968	1546248
31	Yersinia pestis KIM	NP_670101	3103442
31	Yersinia pseudotuberculosis IP 32953	YP_069931	1669768
32	Xenorhabdus nematophila ATCC19061	AAL79611	–
33	Proteus mirabilis	CAA71443	–
34	Photorhabdus luminescens subsp. Laumondii TTO1	NP_928891	1909347
35	Anopheles gambiae str. PEST	XP_306944	–
36	Klebsiella aerogenes W70	AAD12584	–
36	Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	YP_215900	1021167
36	Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC9150	YP_151064	1915298
36	Salmonella typhimurium LT2	NP_459935	1037236
37	Salmonella typhimurium	AAA75467	–
38	Salmonella enterica subsp. enterica serovar Typhi Ty2	NP_805739	2033303
38	Salmonella enterica subsp. enterica serovar Typhi str. CT18	NP_455447	947774
39	Enterobacter aerogenes	AAA75439	–
39	Escherichia coli CFT073	NP_752955	987217
39	Escherichia coli K12 MG1655	NP_415409	932312
39	Escherichia coli O157:H7 EDL933	NP_286766	1153389
39	Escherichia coli O157:H7 RIMD0509952	NP_309001	1064166
39	Shigella flexneri 2a str. 2457T	NP_836547	879273
39	Shigella flexneri 2a str. 301	NP_706774	884460
40	Serratia marcescens	AAA75466	–
41	Klebsiella pneumoniae	AAA75465	–
42	Erwinia carotovora subsp. atroseptica SCRI1043	YP_050739	2981911
43	Photobacterium profundum SS9	YP_129374	I-1303709
44	Vibrio parahaemolyticus RIMD 2210633	NP_797483	I-1162465
45	Vibrio cholerae O1 biovar eltor str. N16961	NP_231538	I-2054880
46	Vibrio vulnificus CMCP6	NP_761755	I-3015110
46	Vibrio vulnificus YJ016	NP_934114	I-1347902
47	Vibrio vulnificus	AAN78126	–
αPBFs No	Species	NCBI code	genome position
48	Caulobacter crescentus CB15	NP_419622	895272
49	Sinorhizobium meliloti 1021	NP_386797	2895144
50	Bradyrhizobium japonicum GX201	AAL35753	–
51	Bradyrhizobium japonicum USDA I110	AAB49303	–
52	Bradyrhizobium japonicum USDA 110	NP_773900	7985928
53	Rubrivivax gelatinosus PM1	ZP_00243330	–
54	Achromobacter denitrificans EST4002	NP_990906	p-33793
55	Polaromonas sp. JS666	ZP_00363432	–
56	Brucella melitensis 16M	NP_539364	I-465504
57	Brucella abortus biovar 1 str. 9-941	YP_222247	I-1540690
57	Brucella suis 1330	NP_698562	I-1522650
TinR No	Species	NCBI code	genome position
94	Salmonella enterica RKS 3333	CAB51578	–
95	Salmonella enterica subsp. enterica serovar Typhi str. CT18	NP_454916	357586
95	Salmonella enterica subsp. enterica serovar Typhi Ty2	NP_806270	2623527
96	Salmonella enterica subsp. enterica serovar Paratypi A str. ATCC9150	YP_151642	2548252
97	Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	YP_215291	352176

Eubacterial FFRPs used as outgroups No	Group	Species	NCBI code	genome position
99	0	Anopheles gambiae str. PEST	XP_306843	–
100	0	Bacteroides fragilis YCH46	YP_101787	5183978
98	5	Mycobacterium tuberculosis H37Rv	NP_217808	3671845

＊: See Archaic Database: http://www.aist.go.jp/RIODB/archaic/index.html/.

AsnC No	Species	NCBI code	genome position
58	Photobacterium profundum SS9	YP_131608	I-3998573
59	Vibrio cholerae O1 biovar eltor str. N16961	NP_229730	I-70975
60	Vibrio vulnificus CMCP6	NP_760059	I-1109653
60	Vibrio vulnificus YJ016	NP_932862	I-80299
61	Vibrio parahaemolyticus RIMD 2210633	NP_796450	I-81432
62	Idiomarina loihiensis L2TR	YP_155084	748704
63	Shewanella oneidensis MR-1	NP_719335	3952103
64	Escherichia coli CFT073	NP_756529	4427578
64	Escherichia coli K12 MG1655	NP_418199	3924568
64	Escherichia coli O157:H7 EDL933	NP_290382	4788970
64	Escherichia coli O157:H7 RIMD0509952	NP_312712	4719989
64	Shigella flexneri 2a str. 2457T	NP_839123	3839481
64	Shigella flexneri 2a str. 301	NP_709556	3933813
65	Salmonella typhimurium LT2	NP_462775	4084699
66	Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	YP_218776	4019754
66	Salmonella enterica subsp. enterica serovar Paratypi A str. ATCC9150	YP_152819	3859718
66	Salmonella enterica subsp. enterica serovar Typhi str. CT18	NP_458067	3764524
66	Salmonella enterica subsp. enterica serovar Typhi Ty2	NP_807280	3750015
67	Yersinia pestis biovar Medievalis str. 91001	NP_991406	554
67	Yersinia pestis CO92	NP_403668	804
67	Yersinia pestis KIM	NP_667347	554
67	Yersinia pseudotuberculosis IP 32953	YP_068552	803
68	Photorhabdus luminescens subsp. laumondii TTO1	NP_927427	45402
69	Erwinia carotovora subsp. atroseptica SCRI1043	YP_048134	565
70	Haemophilus ducreyi 35000HP	NP_874246	1608322
71	Actinobacillus pleuropneumoniae serovar 1 str. 4074	ZP_00135168	–
72	Pasteurella multocida subsp. multocida str. Pm70	NP_246512	1782023
73	Mannheimia succiniciproducens MBEL55E	YP_087227	33356
74	Haemophilus somnus 129PT	ZP_00123006	–
75	Haemophilus somnus 2336	ZP_00132325	–
76	Haemophilus influenzae Rd KW20	NP_438720	582372
76	Haemophilus influenzae R2866	ZP_00156381	–
76	Haemophilus influenzae R2846	ZP_00155555	–
77	Haemophilus influenzae 86-028NP	ZP_00322207	–
YbaO No	Species	NCBI code	genome position
78	Klebsiella pneumoniae CG43	NP_943340	p-46694
79	Salmonella enterica subsp. enterica serovar Paratypi A str. ATCC9150	YP_151460	2356382
79	Salmonella enterica subsp. enterica serovar Typhi str. CT18	NP_455057	509585
79	Salmonella enterica subsp. enterica serovar Typhi Ty2	NP_806130	2471516
79	Salmonella typhimurium LT2	NP_459455	516725
80	Escherichia coli K12 MG1655	NP_414981	468065
81	Escherichia coli O157:H7 RIMD0509952	BAB33924	534904
82	Shigella flexneri 2a str. 2457T	NP_836119	403735
82	Shigella flexneri 2a str. 301	NP_706341	403934
83	Escherichia coli CFT073	NP_752498	543491
84	Escherichia coli O157:H7: EDL933	NP_286189	534907
85	Erwinia carotovora subsp. atroseptica SCRI1043	YP_049263	1312560
86	Photorhabdus luminescens subsp. laumondii TTO1	NP_931065	4535245
87	Yersinia pestis biovar Medievalis str. 91001	NP_992173	848533
87	Yersinia pestis CO92	NP_406621	3506992
87	Yersinia pestis KIM	NP_668368	1170737
87	Yersinia pseudotuberculosis IP 32953	YP_069511	1167117
88	Vibrio cholerae O1 biovar eltor str. N16961	NP_230707	I-1130547
89	Vibrio parahaemolyticus RIMD 2210633	NP_797326	I-986889
90	Vibrio vulnificus CMCP6	NP_761945	I-3228373
91	Vibrio vulnificus YJ016	NP_933927	I-1142509
92	Photobacterium profundum SS9	YP_129183	I-1073657
93	Vibrio fischeri ES114	YP_204961	I-1768877

Archaeal FFRPs used as outgroups Group	Species	Archaic* code	genome* position
1	Pyrococcus sp. OT3	pot0377090	377090
2	Pyrococcus sp. OT3	pot1216151	1216151
2	Pyrococcus sp. OT3	pot0300646	300646
2	Pyrococcus sp. OT3	pot0175330	175330
2	Pyrococcus sp. OT3	pot1664679	1664679
2	Pyrococcus sp. OT3	pot0123002	123002
2	Pyrococcus sp. OT3	pot0301583	301583
2	Pyrococcus sp. OT3	pot0434017	434017
2	Thermoplasma volcanium GSS1	TVG0307586	307586
2	Thermoplasma volcanium GSS1	TVG0651564	651564
2	Thermoplasma volcanium GSS1	TVG1179749	1179749
3	Thermoplasma volcanium GSS1	TVG1409444	1409444
4	Pyrococcus sp. OT3	pot0828564	828564
4	Pyrococcus sp. OT3	pot0836696	836696
4	Pyrococcus sp. OT3	pot0868477	868477
4	Pyrococcus sp. OT3	pot1613368	1613368
4	Pyrococcus sp. OT3	pot1735659	1735659
4	Thermoplasma volcanium GSS1	TVG0368274	368274
4	Thermoplasma volcanium GSS1	TVG1254131	1254131
5	Pyrococcus sp. OT3	pot0008824	8824

Archaeal non-FFRPs used as outgroups Group	Species	Archaic* code	genome* position
4	Pyrococcus sp. OT3	pot0823941	823941
4	Pyrococcus sp. OT3	pot0258936	258936
4	Pyrococcus sp. OT3	pot0112597	112597
4	Thermoplasma volcanium GSS1	TVG1034191	1034191
4	Thermoplasma volcanium GSS1	TVG1497808	1497808
4	Thermoplasma volcanium GSS1	TVG1466300	1466300
4	Thermoplasma volcanium GSS1	TVG0198521	198521

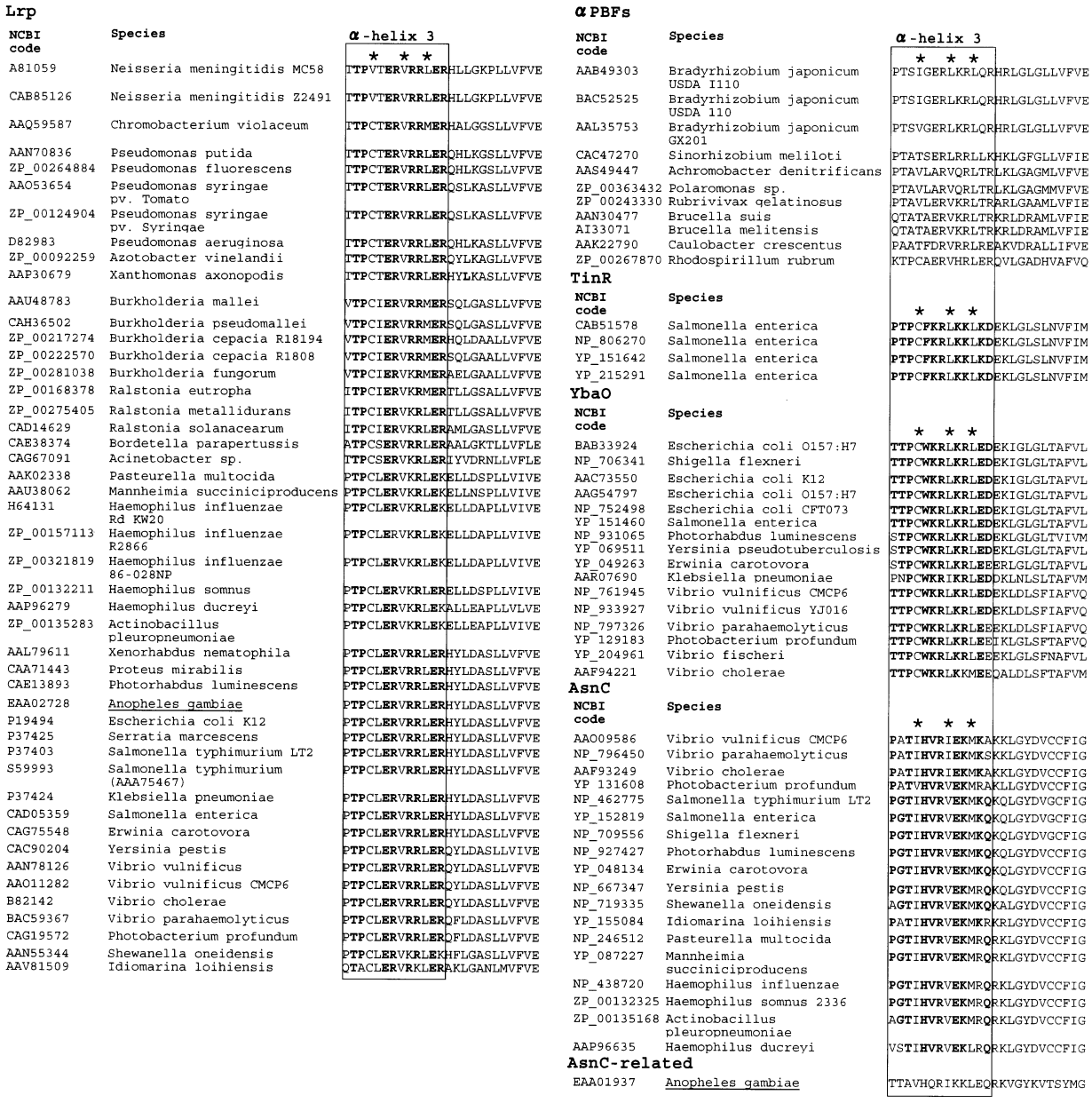

Fig. 1. Comparison of amino acid sequences of FFRPs. Amino acid sequences of α helices 3 (boxed) and twelve positions following the helices. The hydrophobic phases inside the α helices are labeled with "*". Amino acid residues positioned in this phase will interact with other residues inside the protein cores, turning away from the interacting DNA. The remaining positions in the helices will bind DNA phosphates or contact bases. These positions, conserved in each orthologous group, Lrp, TinR, YbaO, or AsnC, are highlighted in bold. Two FFRPs were found coded in DNA fragments extracted with chromosomal DNAs of a mosquito, *Anopheles gambiae* (underlined).

map and its rooted version, i.e. a phylogenetic map, was made by using other types of FFRPs as outgroups (Fig. 2). It is likely that αPBFs have differentiated by direct modifications of Lrp without gene duplication: orthologous non-orthologues.

Notable differences were found between the phylogenetic tree and the standard one in two respects. Firstly, a group of organisms in the β subclass, in the gen-

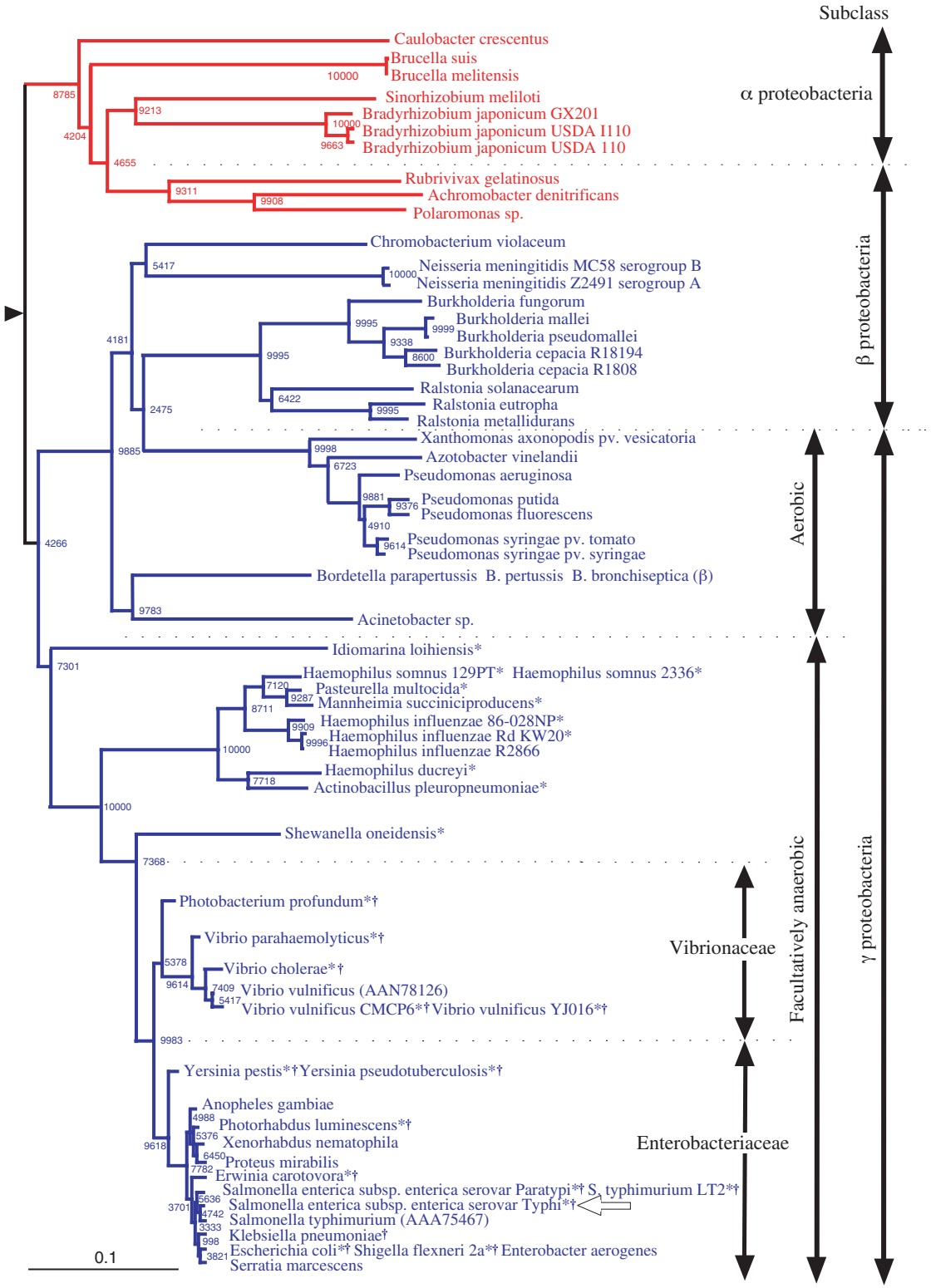

Fig. 2. A distance map made using the amino acid sequences of orthologues of *E. coli* Lrp (blue), and α proteobacterial FFRPs, αPBFs (red). Distribution of orthologues of Lrp and that of αPBFs are complementary to each other. *Proteobacteria* in the facultatively anaerobic order of the γ subclass have another FFRP, AsnC (marked, *). *Vibrionaceae* and *Enterobacteriaceae* have a third FFRP, YbaO (marked, †). A fourth FFRP, TinR, is present in *Salmonella enterica* (the white arrow). Bootstrap values calculated with 10,000 trials are shown for nodes. The root (triangle) was identified using other types of proteobacterial FFRPs.

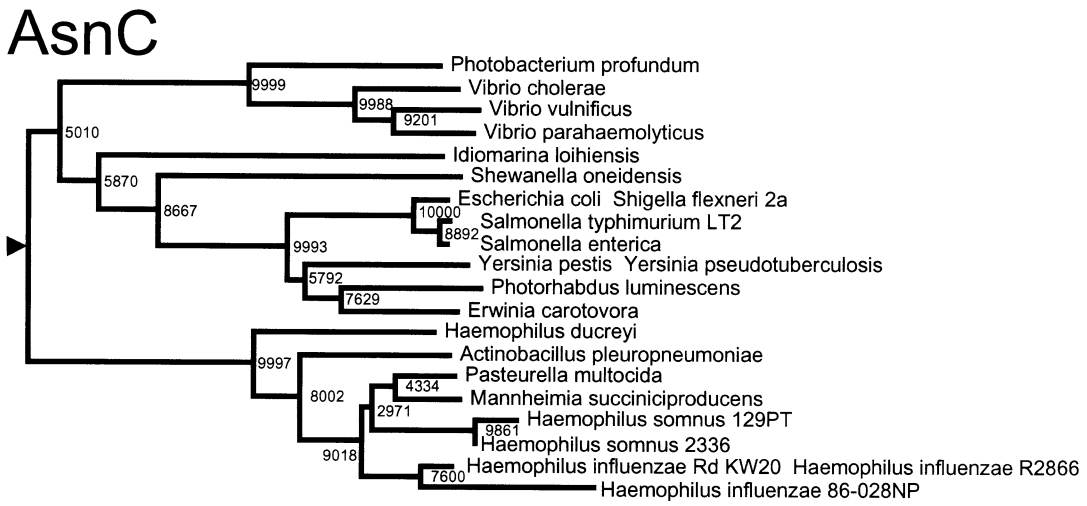

Fig. 3. A distance map made using the amino acid sequences of orthologues of *E. coli* AsnC. The root identified using outgroups, Lrp, YbaO, TinR, is indicated by a triangle. Bootstrap values calculated with 10,000 trials are shown for nodes.

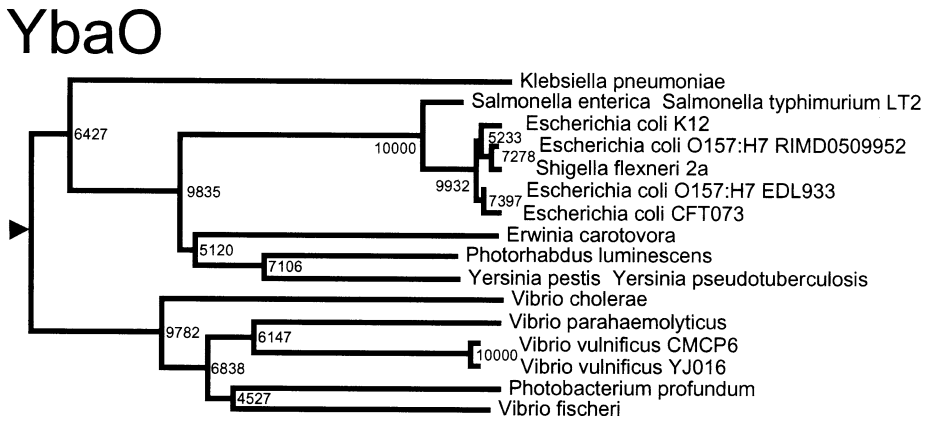

Fig. 4. A distance map made using the amino acid sequences of orthologues of *E. coli* YbaO. The root identified using outgroups, Lrp, AsnC, TinR, is indicated by a triangle. Bootstrap values calculated with 10,000 trials are shown for nodes. The entry *Klebsiella pneumoniae* represents the protein coded in a plasmid.

era *Rubrivivax*, *Achromobacter*, and *Polaromonas*, shared an ancestor with α proteobacteria, differentiating from the rest of the β subclass. In fact, this group had αPBFs instead of Lrp. Secondly, unlike the standard phylogeny, the aerobic order of γ proteobacteria, e.g. *Pseudomonas*, shared a closer ancestor with the β subclass, differentiating from the rest of γ.

The absence of Lrp in the δ and ε subclasses might be due to lack of information: not much is known about these organisms. Some organisms in β and γ, e.g. *Rickettsia*, lacked Lrp, but this fact can be explained by individual losses after speciation. For an organism whose genomic sequence is not determined (marked "-" in Table I in the column of genome position), lack of confirmation of a protein does not necessarily imply its absence. Here the ClustalW program[30] was used for making rooted and unrooted phylogenetic trees.

Orthologues of AsnC were present only in the γ subclass in the order of facultative anaerobes (Fig. 3). It is absent from the order of aerobic γ proteobacteria, suggesting again that this order is closer to β. Among facultatively anaerobic γ proteobacteria, only those in the *Vibrionaciae* and *Enterobacteriaceae* families have the third FFRP YbaO (Fig. 4). Among

Enterobacteriaceae only strains in the species *Salmonella enterica* were identified to have TinR.

Topology of orthologous groups in a distance map. One might imagine that the ordering of the widely distributing Lrp to the most localized TinR could reflect successive gene duplications of *lrp* to *asnC*, of *asnC* to *ybaO etc*. However, this model, here referred to as the "duplication-on-demand" model, is unlikely, as will be discussed in what follows.

When a combined distance map was made by combining the amino acid sequences (Fig. 5), the types of FFRPs clustered separately from each other, so that they became outside to each other (Fig. 6a), thereby filling the center of the map only with the roots of the four orthologous groups (T, A, Y, L in Fig. 6a).

Such a topology is not expected for the "on-demand" model. If gene I could have duplicated and modified to gene II after differentiation of the common ancestor of organisms (CAO_{A+B}) to those of subgroups A and B (CAO_A and CAO_B, respectively), and thus inside CAO_A (Fig. 7b), the root of orthologues II should attach to the stem of subgroup A in cluster I: II being inside the diversification of I, but not outside as observed (Fig. 7a).

In order to explain the observed topology, it is necessary to assume gene duplication inside CAO_{A+B} (Fig. 7c). At this stage the gene was not of I or II but of a common ancestor protein (CAP) of I and II. One of the two copies of CAP differentiated to become orthologue I in CAO_{A+B} (labeled with "1" in Fig. 7c). The other copy remained unused for a number of years, until the CAO differentiated to CAO_A and CAO_B (labeled with "2" in Fig. 7c). Inside CAO_A the unused copy was modified to orthologue II, thereby adopting a new function. In subgroup B this gene was lost, or changed to other proteins. This explanation is referred to as the "duplication-and-wait" model.

By modifying the original "on-demand" model, one might imagine that orthologue II could have been functioning inside CAO_{A+B}, but have been lost in CAO_B due to a change in metabolism or a substitution by another protein having the same function. However, when orthologue II is found in a highly differentiated subgroup only, so that the subgroup is not a primary branch of the group having orthologue I, but a branch of a branch *etc*. it is difficult to assume that after all the differentiation of e.g. *Vibrionaceae* and *Enterobacteriaceae* possessing YbaO, TinR once shared and functioning could have been lost independently except in the single species *Salmonella enterica*.

The "duplication-and-wait" model suggests that a life can be tolerating, and a duplicated protein gene can survive for a number of years without being used. Such tolerance rewards, when the protein adopts a new function. According to the neutral theory an unused gene can change rapidly.[2] It also suggests that duplicated genes found in genomes of extant organisms might not be functioning.

Common ancestor of the four types of FFRPs. An appropriate outgroup, i.e. a protein different from, but close to, the proteobacterial FFRPs, is needed in order to identify the common root of these proteins, thereby converting a distance map to a rooted tree. So far examined, none of the proteobacterial FFRPs, or, in fact, any eubacterial FFRP is present in an archaeon and *vice versa*. Thus FFRPs from two archaea, *Pyrococcus* sp. OT3 and *Thermoplasma volcanium*, were used as outgroups (Table I). For similar reasons, FFRPs from two species in other classes, gram-positive high G/C bacteria (*Mycobacterium tuberculosis*) and bacteroides-flavobacteria (*Bacteroides fragilis*), were used. In addition, seven non-FFRP proteins from the two archaea, having relatively high homology to the proteobacterial FFRPs, were used.

In the distance map (Fig. 5), all the outgroups attached to the connection between AsnC and the rest three types, indicating the common ancestor protein have differentiated to that of AsnC, and CAP of the three other types, from which Lrp, YbaO, and TinR differentiated in this order. For outgroup 5 including the FFRP from *M. tuberculosis* and an FFRP from *Pyrococcus*, the bootstrap value obtained was 5,373 per 10,000 trials. For another FFRP (outgroup 1) a value 4,706/10,000 was obtained. The bootstrap value obtained for the FFRP from *B. fragilis* (outgroup 0 or number 100) was even higher, 8,191/10,000. This protein might not be a real outgroup but might be in a close phylogenetic relation with AsnC.

Two FFRPs coded in DNA fragments isolated with mosquito chromosomes. While we were collecting the amino acid sequences of eubacterial FFRPs, we found that two genes of FFRPs, an *lrp* and another *ffrp* related with *asnC* and the *ffrp* from *B. fragilis* (outgroup 0 or number 100), were coded in DNA fragments sequenced by the company Celera Genomics, while determining the chromosomal sequence of the mosquito *Anopheles gambiae* by the random shot-gun method.[31]

The two DNA fragments were among those remaining unassembled into the chromosomes of the mosquito. No similar sequence was found in the

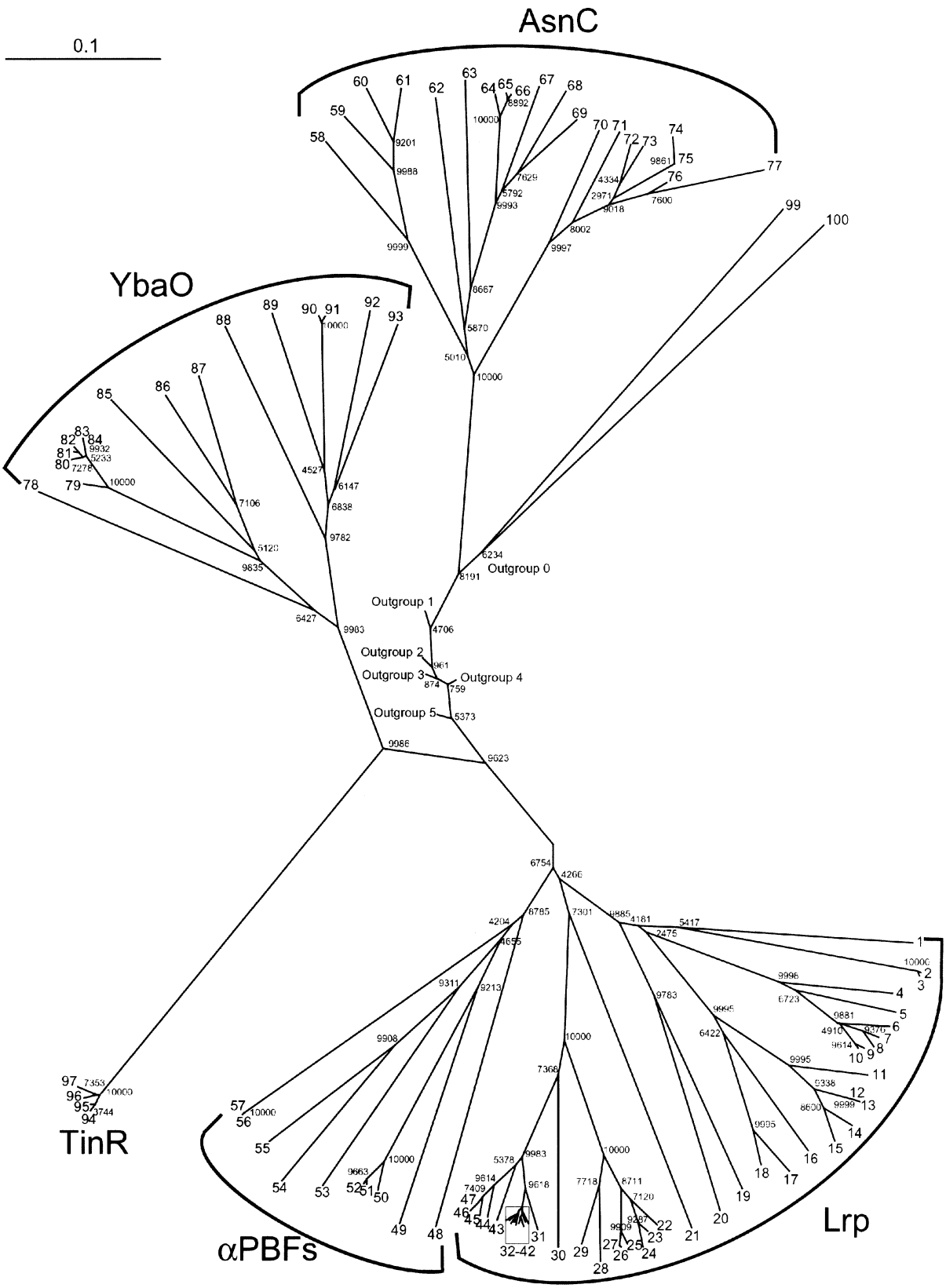

Fig. 5. A combined distance map made using types of FFRPs. See Table I for understanding numbering of FFRPs. Bootstrap values calculated with 10,000 trials are shown for nodes.

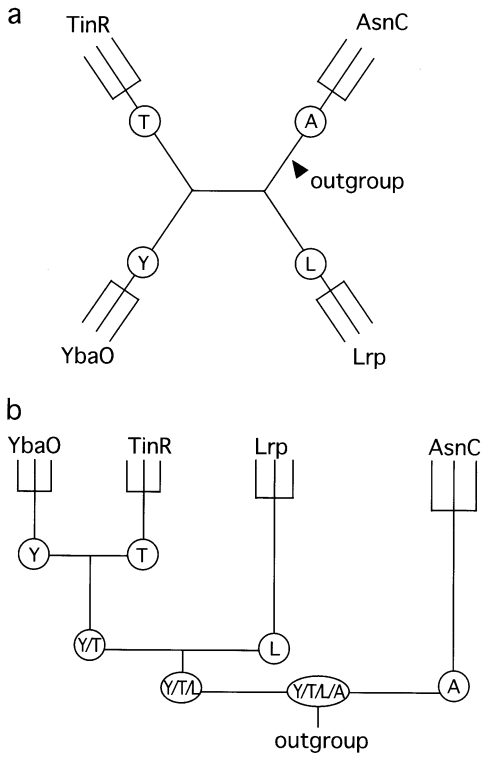

Fig. 6. Schematic representations of the topology in Fig. 5. "L" e.g. indicates the ancestor Lrp. In (b) "Y/T" e.g. indicates the common ancestor of YbaO and TinR. (a) is a simplication of Fig. 5. In (b) the history is traced from the common ancestor "Y/T/L/A" (bottom) to the types of FFRPs (top).

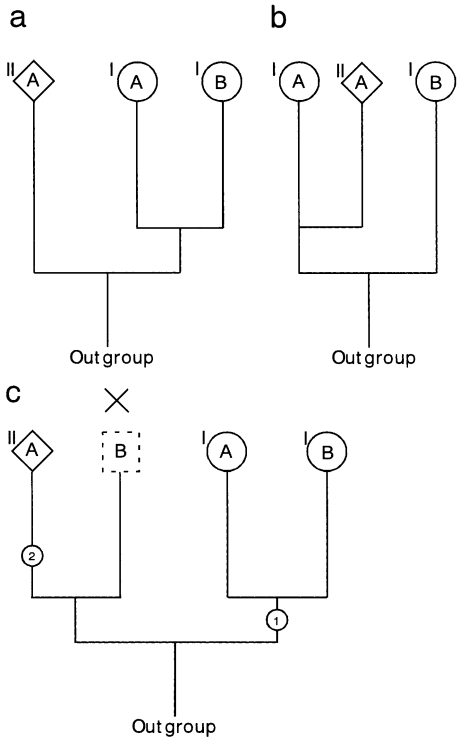

Fig. 7. The observation (a) and the "on-demand" (b) and "wait" (c) models. (a) Any two clusters of FFRPs, here a cluster of circles labeled I and another cluster indicated by a diamond and labeled II, are positioned outside to each other. A and B: groups of organisms. For Lrp (I) and AsnC (II), A is the facultatively anaerobic order and B is the aerobic order of the γ subclass *Proteobacteria*. Outgroups are YbaO and TinR. (b) According to the "on-demand" model, after differentiation of the common ancestor organism to the ancestor of A and that of B, inside the A ancestor the protein gene was duplicated and changed to II. However, this is not consistent with the observation (a). (c) According to the "wait" model, gene was duplicated inside the common ancestor of A and B. "1" indicates the point where orthologue I was created, while "2" indicates the point where orthologue II was created.

genome of *Drosophila melanogaster*,[32] but the Lrp coded in the DNA fragment was found closest to Lrp's from γ enterobacteria in the *Yersina* genus. Of all unassembled DNA fragments of *A. gambiae*, 220 showed high homology to proteins coded in the genome of *Yersina pestis*,[33] at the same level as that of the Lrp (number 35 in Table I and Figs) to *E. coli* Lrp. Of these 41 showed the highest homologies to proteins from organisms in the genus *Yersina*. It is possible that these 125 Kbps cover the same genome of an organism in the genus *Yersina*.

Of all the unassembled DNA fragments sequenced as that of *A. gambiae*, 363 showed high homology to proteins coded in the genome of *Bacteroides fragilis*,[34] at the same level as that of the AsnC-related FFRP (number 99 in Table I and Fig. 5) to the *B. fragilis* FFRP (number 100 in Table I and Fig. 5). Of these 98 showed the highest homology to proteins from organisms in the genus *Bacteroides*. It is possible that these 269 Kbps cover the same genome of another organism in the genus *Bacteroides*.

It is likely that types of eubacteria are living inside the body of the mosquito, and their DNA fragments were co-extracted with the chromosomal DNAs of the mosquito.

Conclusions: two different types of trees. Usually we discuss evolution of organisms by tracing a phylogenetic tree from their CAO (Fig. 8a). Genomic compositions of many extant organism are known, and so variations of proteins they have. Inside each organism, proteins are related by reflecting the process of differ-

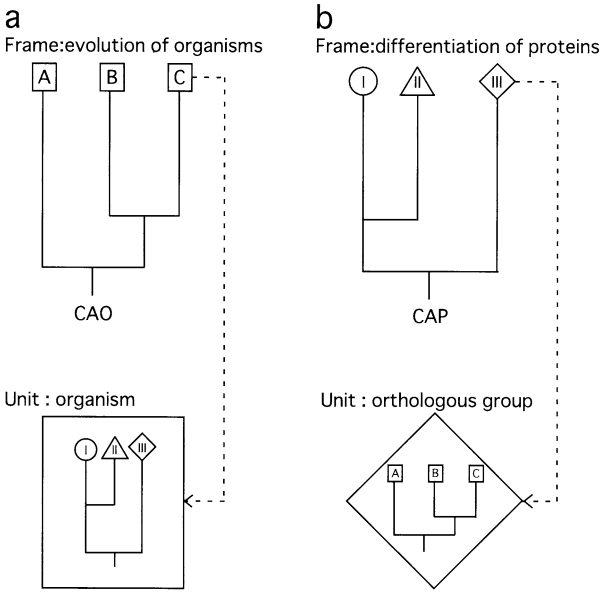

Fig. 8. A structure of biological understanding of evolution of organisms (a) in comparison with that of information obtainable from amino acid sequences of proteins orthologous or paralogous to each other (b). A, B, C: different organisms. I, II, III: orthologous protein groups. CAO: common ancestor organism. CAP: common ancestor protein. In (a) the frame represents the process of evolution of organisms, and in (b) the frame represents the process of differentiation of a protein to paralogues. The unit in (a) is an organism, which has types of proteins related by a topology the same as in the frame of (b). The unit in (b) is a group of orthologous proteins, which are related by a topology the same as that in the frame of (a).

entiation from their CAP. While, in another type of tree such as Fig. 5, inside each group of proteins orthologous to each other the phylogeny of organisms is represented (Fig. 8b), i.e. the frame of an evolution tree (Fig. 8a). The frame of a protein tree (Fig. 8a) represents differentiation of protein types, while such types of proteins can be described in an evolution tree inside units, i.e. organisms. In this way, the two types of trees have representations inside out to each other.

In the past, information obtainable from protein sequences on evolution of organisms have been intensively studied. In the future, information obtainable on differentiation of proteins will be studied more. For this purpose, we need to develop new methods and concepts, since the structure of information analysis does not directly match with the ordinary structure of biology for understanding evolution.

Acknowledgements. This work was supported by the CREST (Core Research for Evolutionary Science and Technology) program of JST (Japan Science and Technology Agency) in the research area Protein Structure and Functional Mechanisms (PSFM).

References

1) Zuckerkandl, E., and Pauling, L. (1962) In Horizons in Biochemistry (eds. Kasha, B., and Pullman, B.). Academic Press, New York, pp. 189-225.
2) Kimura, M. (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.
3) Schmidt-Nielsen, K. (1979) Animal Physiology: Adaptation and Environment. 2nd ed., Cambridge University Press, Cambridge, UK.
4) Patthy, L. (1999) Protein Evolution. Blackwell Science, London.
5) Koike, H., Ishijima, S. A., Clowney, L., and Suzuki, M. (2004) Proc. Natl. Acad. Sci. USA **101**, 2840-2845.
6) Suzuki, M. (2003) Proc. Jpn. Acad., Ser. B **79**, 274-289.
7) Koike, H., Sakuma, M., Mikami, A., Amano, N., and Suzuki, M. (2003) Proc. Jpn. Acad., Ser. B **79**, 63-69.
8) Suzuki, M., Amano, N., and Koike, H. (2003) Proc. Jpn. Acad., Ser. B **79**, 92-98.
9) Suzuki, M., and Koike, H. (2003) Proc. Jpn. Acad., Ser. B **79**, 114-119.
10) Suzuki, M. (2003) Proc. Jpn. Acad., Ser. B **79**, 213-222.
11) Suzuki, M., Aramaki, H., and Koike, H. (2003) Proc. Jpn. Acad., Ser. B **79**, 242-247.
12) Ishijima, S. A., Clowney, L., Koike, H., and Suzuki, M. (2003) Proc. Jpn. Acad., Ser. B **79**, 299-304.
13) Ishijima, S. A., Clowney, L., Koike, H., and Suzuki, M. (2004) Proc. Jpn. Acad., Ser. B **80**, 22-27.
14) Ishijima, S. A., Clowney, L., Koike, H., and Suzuki, M. (2004) Proc. Jpn. Acad., Ser. B **80**, 107-113.
15) Clowney, L., Ishijima, S. A., and Suzuki, M. (2004) Proc. Jpn. Acad., Ser. B **80**, 148-155.
16) Ishijima, S. A., Clowney, L., and Suzuki, M. (2004) Proc. Jpn. Acad., Ser. B **80**, 183-188.
17) Ishijima, S. A., Clowney, L., and Suzuki, M. (2004) Proc. Jpn. Acad., Ser. B **80**, 236-243.
18) Ishijima, S. A., Clowney, L., and Suzuki, M. (2004) Proc. Jpn. Acad., Ser. B **80**, 459-468.
19) Sakuma, M., Koike, H., and Suzuki, M. (2005) Proc. Jpn. Acad., Ser. B **81**, 26-32.
20) Yokoyama, K., Ebihara, S., Kikuchi, T., and Suzuki, M. (2005) Proc. Jpn. Acad., Ser. B **81**, 64-75.
21) Sakuma, M., Nakamura, M., Koike, H., and Suzuki, M. (2005) Proc. Jpn. Acad., Ser. B **81**, 111-117.
22) Kudo, N., Allen, M. D., Koike, H., Katsuya, Y., and Suzuki, M. (2001) Acta Cryst. **D57**, 469-471.
23) Calvo, J. M., and Matthews, R. G. (1994) Microbiol. Rev. **58**, 466-490.
24) Newman, E. B., and Lin, R. (1995) Annu. Rev. Microbiol. **49**, 747-775.
25) Kölling, R., and Lother, H. (1985) J. Bacteriol. **164**, 310-315.

26) Zeng, Q., and Summers, A. O. (1997) Mol. Microbiol. **24**, 231-232.
27) Folkesson, A., Advani, A., Sukupolvi, S., Pfeifer, J. D., Normark, S., and Löfdahl, S. (1999) Mol. Microbiol. **33**, 612-622.
28) Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Nucl. Acids Res. **25**, 3389-3402.
29) Leonard, P. M., Smits, S. H. J., Sedelnikova, S. E., Brinkman, A. B., de Vos, W. M., van der Oost, J., Rice, D. W., and Rafferty, J. B. (2001) EMBO J. **20**, 990-997.
30) Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) Nucl. Acids Res. **22**, 4673-4680.
31) Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M., Wides, R. *et al*. (2002) Science **298**, 129-149.
32) Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F. *et al*. (2000) Science **287**, 2185-2195.
33) Parkhill, J., Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T. G., Prentice, M. B., Sebaihia, M., James, K. D., Churcher, C., Mungall, K. L. *et al*. (2001) Nature **413**, 523-527.
34) Kuwahara, T., Yamashita, A., Hirakawa, H., Nakayama, H., Toh, H., Okada, N., Kuhara, S., Hattori, M., Hayashi, T., and Ohnishi, Y. (2004) Proc. Natl. Acad. Sci. USA **101**, 14919-14924.

(Received April 25, 2005; accepted May 12, 2005)