Tandem Clusters of Membrane Proteins in Complete Genome Sequences

Daisuke Kihara[1] and Minoru Kanehisa[2]

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

The distribution of genes coding for membrane proteins was investigated in 16 complete genomes: 4 archaea, 11 bacteria, and 1 eukaryote. Membrane proteins were identified by our new method of predicting transmembrane segments (Kihara et al. 1998) after the removal of amino-terminal signal peptides. Interestingly, about half of the membrane protein genes in each genome were found to be located next to another, forming tandem clusters. Roughly 10%–30% of the tandem clusters were conserved among organisms, and most of the conserved tandem clusters belonged to one of the three functional groups, namely, transporters, the electron transport system, and cell motility. A tandem cluster sometimes contained paralogous membrane proteins, in which case the cluster size and the number of transmembrane segments could be related to a functional category, especially to transporters. In addition to the clustering of membrane proteins, the clustering of membrane proteins and ATP-binding proteins in the complete genomes was also analyzed. Although this clustering was not statistically significant, it was useful to identify candidate membrane protein partners of isolated ATP-binding protein components in the ABC transporters. Possible implications of tandem cluster organization of membrane protein genes are discussed including the complex formation and other functional coupling of protein products and the mechanism of protein translocation to the cell membrane.

With the wealth of complete genome sequences accumulated by the recent genome projects, we now have the opportunity to analyze genome structure and function comprehensively from the catalog of all the genes encoded in the genome. It is possible, for example, to compare the ordering of genes in different genomes and to understand general principles of how functionally coupled genes are physically encoded in the genome and possibly coregulated at the level of gene expression. The correlation of functional coupling and physical coupling seems to be prevalent in bacterial and archaeal genomes; namely, a set of functionally correlated genes tends to be encoded in a potential operon (Tamames et al. 1997; Dandekar et al. 1998; Overbeek et al. 1999). Thus, the analysis of conserved gene orders among different genomes provides significant clues to functional annotation of individual genes, as additional information to conserved sequence similarity. Furthermore, the prediction of higher order structures may be utilized in order to compensate for the limitation of the sequence similarity search for functional identification. Aurora and Rose (1998) used predicted secondary structures in the search for a particular enzyme and Fetrow and Skolnick (1998) created templates of active sites of enzymes suitable for the screening of genome sequences.

We combine the prediction of membrane proteins with the analysis of gene orders and sequence similarity in the complete genome sequences. Membrane proteins play important roles in living cells, such as for transport, energy production, and cell signaling. Previous studies on membrane proteins in comparative analysis of genome sequences were concerned mostly with the estimation of the number of membrane proteins (Arkin et al. 1997; Boyd et al. 1998; Jones 1998; Wallin and von Heijne 1998). Paulsen et al. (1998) analyzed a specific class of membrane proteins, namely transporters, and discussed the gene distribution in the genome in relation to the environment in which each organism inhabits. Tomii and Kanehisa (1998) also performed a systematic survey of ABC transporters and operon structures in the complete genome sequences, although their analysis did not make use of predictions of transmembrane segments.

Recently, we have developed a new prediction method of membrane proteins, which takes into account the number and the types of transmembrane segments (Kihara et al. 1998). Here, membrane proteins are detected by our prediction method together with the method to remove amino-terminal signal peptides, which are often misidentified as transmembrane segments of mature proteins. It is found that a surprising portion of membrane proteins are encoded as gene clusters, and the pattern of conservation can be used for the prediction of functional categories of membrane proteins. We further report the clustering of membrane proteins and ATP-binding proteins in the genome, which is not statistically significant, but which contains a major class of membrane protein machinery—ABC transporters.

[1] *Current address: Donald Danforth Plant Science Center, St. Louis, Missouri 63141 USA.*
[2] **Corresponding author.**
E-MAIL kanehisa@kuicr.kyoto-u.ac.jp; FAX 81-774-38-3269.

10:731–743 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

RESULTS

Estimated Amounts of Membrane Proteins

First, we analyzed the amount of membrane proteins in the complete genome sequences of 16 organisms. The number of predicted membrane proteins is shown in Table 1. The proportion of membrane proteins in each organism ranges from 18%–29%. This estimate is smaller than that based on a transmembrane prediction method only, which reports the values ~35% (Frishman and Mewes 1997). The discrepancy may be attributable to the removal of amino-terminal signal peptides in our analysis. The estimates for *Haemophilus influenzae* and *Escherichia coli* are almost the same as those by Jones (1998) who also masked out the signal peptides in the preprocessing step; however, his value for *Sacchromyces cerevisiae*, 18%, is much smaller than ours.

The number of probable genes or open reading frames (ORFs) is known to be roughly proportional to the genome size in bacteria and archaea where the gene (ORF) density is about one per 1,000 bases (Table 1). This is to be compared with one per 2,000 bases in *S. cerevisiae* and one per 5,000 bases in *Caenorhabditis elegans*. There is a tendency that the proportion of membrane proteins increases with the number of ORFs in the genome (Wallin and von Heijne 1998); membrane proteins are relatively more abundant in a larger genome, which is also observed in Table 1. Generally speaking, facultative bacteria tend to have larger genomes than obligate bacteria, which are correlated with higher proportions of paralogous genes. Thus, we estimated the proportions of paralogous proteins separately for membrane proteins and nonmembrane proteins. The result is shown in Figure 1. First, paralogs are in fact more abundant in larger genomes, which are clustered in the upper righthand corner of Figure 1A or in the upper portion of Figure 1B. Second, the proportion of paralogous proteins increase with the number of proteins in the pool and seem to become saturated at ~55% (Fig. 1B). Third, when the numbers of proteins are compared in different species, membrane proteins generally contain higher proportions of paralogs than nonmembrane proteins (x vs. ● in Fig. 1B). The additional repertoire of membrane proteins is likely to be used to generate functional diversity and to cope with varying environmental factors.

The distribution of membrane proteins grouped by the number of transmembrane segments was very similar to those reported previously (Arkin et al. 1997; Jones 1998; Wallin and von Heijne 1998; data not shown). Membrane proteins with three or more predicted transmembrane segments are highly likely to be true membrane proteins (see Discussion). They constitute roughly half of all the membrane proteins in our analysis (36%–60% in Table 1).

The Tandem Clusters of Membrane Proteins

Next, we investigated the gene distribution pattern of membrane proteins in the complete genome sequences. Surprisingly, in all 16 organisms ~50%of the membrane proteins were found to be located next to each other, namely, in tandem clusters. Table 2 shows the number of membrane proteins in tandem clusters, as well as the number of tandem clusters of membrane proteins in each genome. The statistical significance of the number of membrane proteins in tandem clusters, which we call here the score, was estimated by randomizing the locations of all the ORFs in the genome.

Table 1. The Numbers of Predicted Membrane Proteins and ATP-Binding Proteins

Category	Organism (abbreviation)	Genome size	Total ORFs	Membrane protein (%)[a]	Mem. prot. with >3 TM segs (%)[b]	ATP-binding proteins (%)[a]
Archaea	*M. jannaschii* (Mj)	1,664,987	1,735	326 (18.8)	154 (47.2)	130 (7.5)
	M. thermoautotrophicum (Mt)	1,751,377	1,871	395 (21.1)	201 (50.9)	127 (6.8)
	A. fulgidus (Af)	2,178,400	2,407	499 (20.7)	268 (53.7)	158 (6.6)
	P. horikoshii (Ph)	1,738,505	1,829	433 (23.7)	235 (54.3)	148 (8.1)
Bacteria	*E. coli* (Ec)	4,639,221	4,289	1142 (26.6)	621 (54.4)	293 (6.8)
	H. influenzae (Hi)	1,830,135	1,717	378 (22.0)	217 (57.4)	153 (8.9)
	H. pylori (Hp)	1,667,867	1,566	334 (21.3)	161 (48.2)	118 (7.5)
	B. subtilis (Bs)	4,214,814	4,100	1125 (27.4)	659 (58.6)	277 (6.8)
	M. genitalium (Mg)	580,073	467	90 (19.3)	54 (60.0)	61 (13.1)
	M. pneumoniae (Mp)	816,394	677	123 (18.2)	72 (58.5)	73 (10.8)
	M. tuberculosis (Mtu)	4,411,529	3,918	949 (24.2)	451 (47.5)	213 (5.4)
	B. burgdorferi (Bb)	910,724	1,256	297 (23.6)	117 (39.4)	85 (6.8)
	T. pallidum (Tp)	1,138,011	1,031	228 (22.1)	120 (52.6)	95 (9.2)
	Synechocystis sp. (Ss)	3,573,470	3,166	918 (29.0)	362 (39.4)	218 (6.9)
	A. aeolicus (Aa)	1,551,335	1,522	304 (20.0)	171 (56.3)	130 (8.5)
Eukarya	*S. cerevisiae* (Sc)	12,069,313	6,215	1652 (26.6)	596 (36.1)	414 (6.7)

[a] Percentage of the total number of ORFs.
[b] Percentage of the total number of membrane proteins.

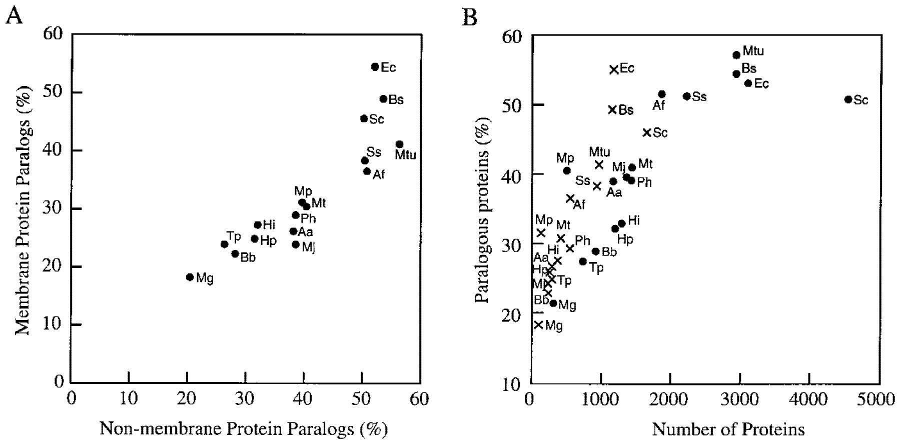

Figure 1 (*A*) The proportion of paralogous proteins is plotted for membrane proteins vs. nonmembrane proteins in 16 organisms (see Table 1 for abbreviations). A paralog is defined by the Smith-Waterman score of 150 or more by SSEARCH after preprocessing with SEG. (*B*) The proportion of paralogous proteins is plotted against the total number of membrane or nonmembrane proteins in 16 organisms. (x) Membrane proteins, (●) nonmembrane proteins.

For each organism, the randomization was performed 1000 times and the mean and the standard deviation of the scores were calculated. The actual score can then be converted to the Z value (Table 2), which is the number of standard deviation units from the mean. Assuming that the score follows the normal distribution, the probability of observing the actual score by chance alone can be determined. The probability was <0.03%, except for *Treponema pallidum* (4.6%) and *Pyrococcus horikoshii* (2.7%).

We have defined a tandem cluster simply as a group of adjacent membrane protein genes in the genome. However, in most of the cases a tandem cluster is formed by the genes encoded on the same strand (Table 2, fourth column) and, furthermore, the gaps between genes are usually <300 bp, which is the condition used by Overbeek et al. (1999) to define a gene cluster (Table 2, fifth column). Thus, it is possible that most of the tandem clusters correspond to operon structures, except for *S. cerevisiae*. In ⅕ of the cases, a tandem cluster spans both strands although it is not clear whether it is under the same gene regulatory mechanism. About 60%–80 % of the tandem clusters are of size two. The longest cluster is found in *Methanobacterium thermoautotrophicum*, containing 12 genes on the same strand, MTH384–MTH395, which corresponds to 10 functionally unknown proteins and 2 subunits of NADH dehydrogenase.

Conserved Tandem Clusters

When the sequence similarity of constituent membrane proteins was examined, some of the tandem clusters were conserved between organisms and/or within an organism. The proportion of such conserved tandem clusters is shown in the last column of Table 2. It was in the range of 10%–30% in bacteria and archaea, but it was 2% in *S. cerevisiae*. The conserved clusters almost exclusively (97.6%) consisted of membrane protein genes encoded on the same strand with gaps of <300 bp, most likely representing conserved operons.

The majority (97.3%) of the conserved tandem clusters could be associated with known functions, which are summarized in Table 3. They belong to one of the three functional categories: membrane transporters, the electron transport system, and cell motility. A large fraction of the transporter category was formed by the ABC (ATP-binding cassette) transporters (Higgins 1992; Fath and Kolter 1993; Dean and Allikmets 1995). The conserved operon structures of ABC transporters are known to be related to the grouping of substrate specificity (Saurin and Daussa 1994; Tomii and Kanehisa 1998), which is also observed in Table 3. The average cluster size for the transporters is small, ~2.6. In contrast, the electron transport system consists of a larger cluster size, ~5, and the constituent membrane proteins are also larger, with >10 transmembrane segments. In many cases the membrane proteins encoded in a conserved tandem cluster, i.e., in a conserved operon, are likely to interact physically (Dandekar et al. 1998)— two permease proteins forming a channel for the ABC transporter, multiple subunits forming an enzyme complex, or multiple subunits responsible for chemotaxis and flagellar assembly. In the last category of other functions in Table 3, rod shape-determinant protein (RodA) and penicillin-binding protein 2 (Pbp2) are responsible for the cell wall formation; RodA activates Pbp2, which synthesizes peptideglycan.

In addition to the conserved clusters shown in Table 3, there were eight conserved clusters consisting of hypothetical membrane proteins as shown in Table 4. Two of them are conserved among three organisms; cluster no. 1 is conserved in *E. coli*, *H. influenzae*, and *Bacillus subtilis* and cluster no. 2 is conserved in *Methanococcus jannaschii, Archaeoglobus fulgidus*, and *P. horikoshii*.

As mentioned above, *S. cerevisiae* has a very low rate of conservation of tandem clusters, and all the eight conserved clusters of *S. cerevisiae* (Table 2) are conserved only within the organism. This is consistent with the fact that *S. cerevisiae* doesn't have bacteria-like operons (Zhang and Smith 1998). For example, the

Table 2. The Number of Membrane Proteins in Tandem Clusters

	Number of membrane proteins in tandem clusters				Number of tandem clusters					
					size of tandem clusters					
Organism	total (%)[a]	Z-value	same strand (%)[b]	same strand and gaps <300 bp (%)[b]	2	3	4	5	>5	conserved (%)[c]
M. jannaschii	160 (49.1)	4.63	119 (74.3)	112 (70.0)	51	10	4	1	1	9 (13.4)
M. thermoautotrophicum	215 (54.4)	5.65	159 (74.0)	156 (72.6)	48	14	11	2	2	7 (9.1)
A. fulgidus	268 (53.7)	6.25	229 (85.4)	227 (84.7)	81	19	5	2	3	10 (9.1)
P. horikoshii	201 (46.4)	1.93	149 (74.1)	140 (69.7)	52	19	7	1	1	17 (17.7)
E. coli	634 (55.5)	5.49	490 (77.3)	456 (71.9)	179	42	20	5	7	56 (22.1)
H. influenzae	198 (52.5)	4.52	165 (83.3)	153 (77.3)	57	12	8	1	1	25 (31.3)
H. pylori	176 (52.7)	4.58	150 (85.2)	144 (81.8)	50	16	3	2	1	10 (13.9)
B. subtilis	670 (59.6)	6.93	500 (74.6)	467 (69.7)	192	51	12	4	10	46 (17.1)
M. genitalium	47 (52.2)	3.47	42 (89.4)	36 (76.6)	14	5	1	0	0	6 (30.0)
M. pneumoniae	65 (52.9)	3.74	56 (86.2)	52 (80.0)	19	6	1	1	0	6 (22.2)
M. tuberculosis	504 (53.2)	5.63	363 (72.0)	352 (69.8)	138	46	9	4	5	36 (17.8)
B. burgdorferi	137 (60.1)	5.41	109 (79.6)	107 (78.1)	30	14	3	2	2	10 (19.6)
T. pallidum	100 (43.9)	1.68	79 (79.0)	79 (79.0)	30	10	1	0	1	11 (26.2)
Synechocystis sp.	512 (55.8)	3.63	409 (80.0)	379 (74.0)	129	35	20	3	7	26 (13.4)
A. aeolicus	141 (46.4)	3.41	109 (77.3)	106 (75.2)	46	4	4	0	3	10 (17.5)
S. cerevisiae	858 (51.9)	4.09	507 (59.1)	190 (22.1)	255	69	26	2	4	8 (2.2)

[a] Percentage to the total number of membrane proteins shown in Table 1.
[b] Percentage to the total number of membrane proteins in tandem clusters.
[c] Percentage to the total number of tandem clusters.

genes for the subunits of cytochrome c oxidase and ATP synthase are scattered in different chromosomes.

Gene Duplication in Tandem Clusters

In the above analysis of conserved clusters, the sequence similarity of membrane proteins was used to define the similarity relationship between two tandem cluster units. Here the sequence similarity is examined within a tandem cluster unit to identify possible gene duplications of constituent membrane proteins. More than 10% of the tandem clusters in each organism are found to contain paralogs, i.e., pairs of constituent membrane proteins that are similar to each other. In Figure 2 the number of such membrane proteins with similar partners is compiled against the size of the belonging clusters and the predicted number of transmembrane segments. To correlate with functional information, the compilation is made separately for three groups: ABC transporters, other transporters, and the rest. The membrane proteins of ABC transporters (Fig. 2A) exhibit the most characteristic features. The cluster size is two in most cases and the number of transmembrane segments peaks at around six and seven (Higgins 1992; Tam and Saier 1993). At the same time there are significant variations of both the number of transmembrane segments and the number of membrane proteins in an operon. The maltose transporter (MalF) is experimentally known to contain eight transmembrane segments (Froshauer et al. 1988), but the number of predicted transmembrane segments can vary more drastically as seen in Figure 2A.

The membrane proteins for the transporters other than the ABC transporters are shown in Figure 2B. The cluster sizes of cation ATPases are three in most cases. Membrane proteins with 10 predicted transmembrane segments and in the cluster of size three are sodium- and calcium-transport ATPases. Clusters of larger sizes (>3) are a mixture of membrane proteins of various kinds, though some of them are still hypothetical proteins and thus their functions could not be assigned. Membrane proteins with clusters of other functions (Fig. 2C) are clearly differentiated into two groups: those with a larger cluster size (~6) and many transmembrane segments (~12) and those with a smaller size (~3) and a few transmembrane segments (1 or 2). The former group corresponds to the complexes of membrane proteins in the respiratory chain, namely, NADH dehydrogenase and cytochrome c oxidase. The latter group corresponds to various kinds of membrane proteins, including peptide synthase, kinase, methyl-accepting chemotaxis protein, surfactin synthase, and others.

The observation made here may be used for functional assignment of membrane proteins without any sequence similarity to known proteins. In fact, a simple rule of discriminating transporters can be established according to the frequencies of transporters and non-transporters. When a tandem cluster of membrane proteins contains paralogs, and if the number of predicted transmembrane segments and the size of the cluster are in the range shown in Figure 3, then the cluster is likely to be a transporter. A darker box in Figure 3 represents

Table 3. Conserved Tandem Clusters with Known Functions

Category	Function	Average cluster size	Organisms (Number of clusters[a])	Typical arrangement of membrane proteins[b]
Membrane transport	ABC transporter (sugar, phosphate, sulfate, spermidine/putrescine, sn-glycerol-3-phospate)	2.2	Mj, Mt, Af (2), Ph (4), Ec (7), Hi (2), Bs (8), Mg (2), Mp (2), Mtu (7), Bb (2), Tp (2), Aa, Ss (4)	Ec: b1124[6]–b1125[6]
	ABC transporter (di-, oligo-peptide)	2.1	Af, Ph (4), Ec (7), Hi (3), Hp (2), Bs (3), Mg, Mp, Mtu (2), Bb (2)	Ec: b1485[6]–b1486[6]
	ABC transporter (ribose)	2.7	Ph, Bs, Mg, Mp, Bb (2), Tp (3)	Mp: A65_orf517[9]–A65_orf311[7]
	ABC transporter (branched-chain amino acid, sugar)	2.1	Mj, Af (4), Ec (2), Ss	Ec: b3456[8]–b3457[7]
	ABC transporter (polar amino acid)	2.1	Ec (4), Hi, Hp, Bs (2)	Bs: glnP[4]–glnM[6]
	ABC transporter (ferrichrome, iron)	3.4	Ec (2), Bs (5), Ss	Bs: fhuG[9]–fhuB[10]
	ABC transporter (iron (III) dicitrate)	2.0	Hi, Bs, Tp	Bs: ytgD[9]–ytgC[7]
	ABC transporter	2.8	Ph, Ec (2), Hi (2), Bs (3), Mg, Mp, Mtu (3)	Ec: b0886[4]–b0887[6]
	ABC transporter?	2.5	Mtu (4)	Mtu: Rv0587[6]–Rv0588[6]
	PTS system IID, IIC component	2.7	Ec (2), Bs	Ec: b1818[7]–b1819[4]
	Cobalt transport permease and cobalamin biosynthesis protein	3.3	Mj (2), Mt (2), Af (2), Ss	Mj: MJ1089[5]–MJ1090[2]–MJ1091[6]
	Branched-chain amino acid transporter	3.0	Hi, Hp, Bs	Bs: azlD[4]–azlC[6]
	Glucose, hexose transporter	3.0	Sc (3)	Sc: YDR342[9]–YDR343C[9]
	Protein export membrane protein	2.9	Ph, Ec, Hi, Hp, Mtu, Bb, Tp, Ss	Hp: HP1549[5]–HP1550[5]
	Biopolymer transport protein	3.3	Ec (2), Hi (2), Hp (2), Ss (3)	Ec: b3005[1]–b3006[2]
	Multidrug resistance protein Major facilitator family transporter	2.5	Ec (3), hi, Aa, Bs	Ec: b2685[1]–b2686[10]
	ATP synthesis (A, B, C chain)[c]	4.8	Ec, Hi, Bs, Mtu, Ss	Bs: atpF[1]–atpE[1]–atpB[5]
	K^{+} transporting ATPase (A, B, C chain)	3.5	Ec, Mtu, Ss	Ss: slr1728[9]–slr1729[7]–slr1730[1]
	Transporting ATPase (Na^{+}, Ca^{2+}, H^{+})	3.0	Sc (2)	Sc: YDR038C[10]–YDR039C[10]–YDR040C[10]
Electron transport system	NADH (monoxide) dehydrogenase quinone oxidoreductase formate hydrogenase, hydrogenase-4	5.5	Af, Ph, Ec (3), Hp, Bs, Mtu (2), Aa (2), Ss (3)	Ec: b2276[11]–b2277[12]–b2278[13]–b2279[1]–b2280[3]–b2281[1]–b2282[7]
	Cytochrome c oxidase cytochrome o ubiquinol oxidase	5.6	Aa (2), Bs (2), Ec, Ss (3), Mtu	Bs: ctaA[7]–ctaB[7]–ctaC[3]–ctaD[14]–ctaE[4]
	Cytochrome d ubiquinol oxidase cytochrome bd II oxidase	4.0	Ec (2), Hi, Bs, Mtu, Aa, Ss	Ec: b0733[6]–b0734[7]
	Cytochrome c-type biogenesis protein	5.0	Ec, Hi	Ec: b2196[13]–b2197[1]–b2198[1]–b2199[6]–b2200[7]
Cell motility	Flagellar motor protein chemotaxis motB,A protein	2.3	Ec, Hp, Bs (2), Bb, Tp, Aa	Ec: b1889[1]–b1890[4]
	Flagellar biosynthesis protein	5	Ec (2), Bs, Tp, Aa	Bs: fliP[4]–fliQ[2]–fliR[6]–flhB[4]–flhA[7]
	Metyl-accepting chemotaxis protein	2.9	Ec, Bs, Bb (2), Tp, Ss (2)	Bb: BB0596[2]–BB0597[2]
	Spore germination protein	2.0	Bs (3)	Bs: yndD[5]–yndE[10]
Others	Two-component system (sensor kinase, response regulator)	2.2	Ec (2), Bs, Ss	Ec: b2218[1]–b229[1]
	Formate dehydrogenase	2.0	Ec (2), Hi, Aa	Ec: b1475[1]–b1476[4]
	NAD(P) transhydrogenase	3.3	E, Hi, Mtu	Ec: b1602[7]–b1603[5]
	Rod shape-determinant protein and penicillin binding protein 2	2.5	Ec, Hi, Bb, Tp	Ec: b0634[8]–b0635[1]
	HflK, HflC γCII stability-governing protein	2.7	Ec, Hi, Tp	Ec: b4174[1]–b4175[1]
	Tetrahydromethanopterin S-methyltransferase	4	Mj, Mt	Mj: MJ0847[5]–MJ0848[6]–MJ0849[8]–MJ0850[1]

[a] The number is 1 if not specified.
[b] The predicted number of transmembrane segments in brackets.
[c] ATP synthase and ATPase are also categorized electron transport system.

Table 4. Conserved Tandem Clusters with Unknown Functions

Cluster number	Organism	Conserved protein 1[a]	TM[b]	Conserved protein 2	TM	Conserved protein 3	TM	Cluster size	Note
1	Ec	b2141 (+)	4	b2142 (+)	4			2	
	Hi	HI1297 (+)	4	HI1298 (+)	5			2	
	Bs	ywbH (+)	4	ywbG (+)	5			3	
	Bs	ysbA (+)	5	ysbB (+)	3			2	
2	Mj	MJ1561 (−)	1	MJ1562 (−)	5			3	
	Af	AF1230 (−)	1	AF1229 (−)	9			2	
	Ph	PH0286 (−)	1	PH0287 (−)	13			2	
3	Ec	b3577 (+)	4	b3578 (+)	10			3	
	Hi	HI1030 (−)	3	HI1029 (−)	13			2	
	Hi	HI0051 (−)	4	HI0050 (−)	9			4	
4	Mg	MG225 (+)	10	MG226 (+)	11			2	
	Mp	F10_orf491 (−)	10	F10_orf503 (−)	12			2	
5	Mg	Mg241 (+)	3	Mg242 (−)	2			3	
	Mp	F10_orf621 (+)	3	F10_orf632o (−)	3			3	
6	Mtu	Rv0450c (−)	12	Rv0451c (−)	1			3	Protein 3 is long-chain-fatty-acid CoA ligase
	Mtu	Rv0402c (−)	12	Rv0403c (−)	1	Rv0404 (+)	1	4	
	Mtu	Rv1522c (−)	12			Rv1521c (−)	3	6	
	Mtu	Rv2942 (−)	12			Rv2941 (+)	2	2	
7	Mtu	Rv0283 (+)	1	Rv0284 (+)	1			2	Protein 2 is similar to cell division proteins
	Mtu	Rv1782 (+)	1	Rv1783 (+)	2			2	
	Mtu	Rv3869 (+)	1	Rv3870 (+)	3			3	
	Mtu	Rv3895c (−)	2	Rv3894c (−)	2			2	
8	Mtu	Rv1796 (+)	2	Rv1795 (+)	9	Rv1797 (+)	1	3	
	Mtu	Rv3886c (−)	2	Rv3887c (−)	11	Rv3885c (−)	1	3	
	Mtu	Rv3876 (+)	1	Rv3877 (+)	11			2	

[a] The strand direction shown in parentheses.
[b] The number of predicted transmembrane segments.

higher likelihood, which is defined in three levels according to the relative frequency and the absolute number of transporters in Figure 2. Level 1 corresponds to the relative frequency of 0.85 or higher with ⩾12 observed instances. Levels 2 and 3 correspond to the relative frequency of 0.8 or higher with ⩾5 and 3 instances, respectively. Although this empirical rule is derived from a small number of samples, we believe that it is still useful to obtain any functional clue to the large number of genes left unassigned in the completely sequenced genomes. In our data set there were 55 tandem clusters containing paralogs whose functions were not known. Based on this empirical rule we predict seven transporters which are shown in Table 5.

Clustering of Membrane Proteins and ATP-binding Proteins

Because ATP-binding proteins provide energy for active membrane transport and other cellular machineries, we suspected that there would be a tendency for membrane proteins and ATP-binding proteins to form clusters in the genome. This is certainly the case for the ABC transporters, whose operons generally contain adjacent permease proteins and ATP-binding proteins. In the 16 complete genomes, ATP-binding proteins constituted between 5% and 13% of ORFs (Table 1). We defined a cluster for each ATP-binding protein together with all the membrane proteins and ATP-binding proteins within five gene positions on both sides. When such physical coupling of membrane proteins and ATP-binding proteins was searched, roughly 21%–39% of ATP-binding proteins were adjacent to membrane proteins. However, this coupling was not statistically significant (data not shown). On average, ~⅓of all the pairs of membrane proteins and ATP-binding proteins are conserved among different organisms as shown in Table 6. More than one-half of the conserved pairs fall in the category of ABC transporters. In addition, the conserved pairs include protein-export proteins, proteins involved in twitching motility, flagellar biosynthesis or secretion, pairs of gluconokinase and gluconate transporter, V-type ATP synthases, polyketide synthases, acriflavin resistance proteins, sporulation proteins, pairs of signal recognition particle protein, and protein-export membrane protein. Clusters of unknown functions were also detected, but they were conserved only between two closely related species.

There was a relatively large group of conserved pairs within *S. cerevisiae*. This paralog group consists of five adjacent pairs, namely, (YML133C, YML132W), (YNL339C, YNL336W), (YHL050C, YHL048W), (YGR296W, YGR295C), and (YFL066C, YFL062W), respectively for the ATP-binding protein and the membrane protein. The membrane proteins are similar to

A

Number of TM Segments

Cluster Size

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2		1	1	8	15	38	47	21	6	12	3						
3				3	4	14	3	2	1								
4					3	5	4	1									
5					1		1	1									
6								1									
7																	
8																	
9																	
10																	
11																	

B

Number of TM Segments

Cluster Size

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2				2	9	4	2	1	6	5	4	5	1		1		
3	2				2	3	1		1	9	3						
4		2	1	2	1	1						2					
5	1		1		3					1							
6																	
7																	
8	2					2	1		1								
9																	
10																	
11																	

C

Number of TM Segments

Cluster Size

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2	19	7	3	7	3	4	2	2	2		1	1		1	1		
3	11	2			1	1	1		1	7		1				1	
4	19	4	2	1				1	2						1		
5		13	2	1	1	1				2	2	4	2	2			
6											2	2	7	2		2	1
7								2		1	6	2	3				
8	1	1									1		1				
9																	
10													2				
11																	

Figure 2 The frequency of membrane proteins that are paralogous within a tandem cluster, where each membrane protein is classified according to the number of transmembrane segments and the size of the cluster to which it belongs. The criterion for a paralog is the same as in Figure 1. The frequency is counted separately for three functional groups: (*A*) ABC transporters; (*B*) transporters other than ABC transporters; and (*C*) membrane proteins of other functions. In total, 256 clusters (13.0%) contained such pairs of paralogous membrane proteins, and among them, 98 (38.3%) were those of ABC transporters, 37 (14.5%) were those of other transporters, 63 (24.6%) were those of the other functions, and 58 (22.7%) were those of hypothetical functions (not shown).

subtelomerically encoded proteins and are predicted to have three transmembrane segments, except for YGR295C that contains just two. We believe that the physical coupling of these membrane proteins and the ATP-binding proteins has some functional relevance.

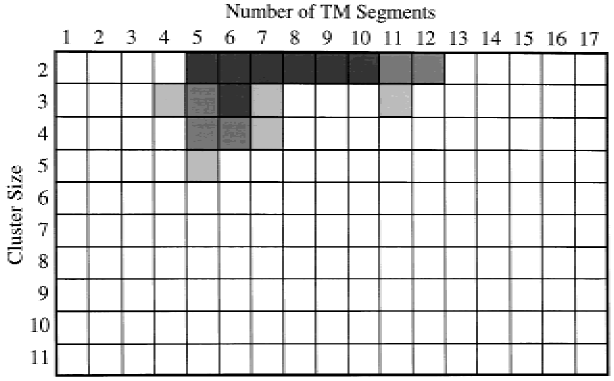

Figure 3 When a tandem cluster of membrane proteins contains paralogs, the cluster size and the numbers of transmembrane segments in paralogous membrane proteins may be used to assign functions. Here the likelihood of being a transporter (*A* or *B* in Fig. 2) rather than in another functional category (*C* in Fig. 2) is shown by the darkness in three levels, darker meaning more likely. The likelihood is defined by the relative frequency and the absolute number of transporters in Figure 2 (see text for details).

Prediction of ABC Transporter Components

The ABC transporters form the largest superfamily of paralogous proteins in bacterial and archaeal genomes (Tatusov et al. 1997; Paulsen et al. 1998). Typically, a transporter consists of three components: a pair of ATP-binding proteins, a pair of membrane proteins, and a substrate-binding protein. In bacteria and archaea the majority of these components are known to be located next to each other, probably forming operons (Tomii and Kanehisa 1998), but there are also isolated components. We have searched by sequence similarity such isolated components of ABC transporters in the complete genomes (Table 7) and tried to identify their partners (Table 8). An ABC-transporter component was considered to be isolated when there was no other component within five gene positions on both sides. Note that the search was performed using the annotated set of bacteria-type ABC transporters in KEGG (Tomii and Kanehisa 1998). Therefore, many eukaryotic ABC transporters (Fath and Kolter 1993) in *S. cerevisiae* were not detected.

As shown in Table 7, the degree of isolation depends on the organism; in *Synechosystis* and *Aquifex aeolicus* ~40% of the components are isolated. Table 7 also indicates that ATP-binding protein components are more likely to be isolated than membrane protein components. However, this may simply be due to the fact that because membrane protein components are less conserved than ATP-binding protein components (Tomii and Kanehisa 1998), they have not been detected by sequence similarity searches. Candidates of missing membrane protein components, which are the partners of isolated ATP-binding protein components, may then be found by examining conserved pairs of ATP-binding proteins and membrane proteins. Table 8

Table 5. Predicted Transporters According to the Number of Transmembrane Segments and the Cluster Size

Organism	Homologous proteins[a]	TM	Additional protein	TM	Cluster size	Reliability[b]	Note
Mj	MJ0419 (+)	10			2	Level 1	Near to MJ0423 which has an ATP binding motif
	MJ0420 (+)	9					
Ec	b0786 (+)	7	b0788 (−)	7	3	Level 3	Similar to Ctr: CT819 (transport permease)
	b0787 (+)	7					
Bs	YybM (+)	5	yybL (+)	5	3	Level 2	Next to yybJ which is an ABC transporter ATP-binding protein
	YybK (+)	5					
Mg	MG225 (+)	10			2	Level 2	Similar to Mt: MTH546 (cationic amino acid transporter related protein)
	MG226 (+)	11					
Bb	BB0050 (+)	5			2	Level 1	
	BB0051 (+)	5					
Bb	BB0807 (+)	5	BB0806 (+)	1	3	Level 2	
	BB0808 (+)	6					
Tp	TP0883 (−)	6			2	Level 1	
	TP0884 (−)	5					

[a] The strand direction shown in parentheses.
[b] The reliability level is according to Figure 3 and based on the worst level for different numbers of transmembrane segments.

summarizes the results of searching for missing membrane protein components. The newly identified pairs of the ATP-binding protein and membrane protein components are predicted to form new types of ABC transporters. Note that the maximum distance of two genes in a cluster is 10 gene positions, which is the case for HI1252 and HI1242 in cluster no. 1 in Table 8. They are in the same cluster that has HI1247 as its center, and because HI1247 is not a component of an ABC transporter, both HI1252 and HI1242 were termed to be isolated in Table 7.

Table 7 also shows the number of fused components. The majority of fused components belong to the class of multidrug-resistance family transporters (Tomii and Kanehisa 1998) where the membrane protein component is fused with the ATP-binding protein component. Occasionally two fused components are encoded as tandem repeats in the genome. In *S. cerevisiae*, most of the components are fused and isolated, but bacteria-type single domain components are also found.

Table 6. The Number of Adjacent Pairs of Membrane Proteins and ATP-binding Proteins

	Number of membrane protein/ATP-binding protein pairs		
Organism	total (%)[a]	same strand (%)[b]	conserved (%)[b]
M. jannaschii	27 (20.8)	22 (81.5)	5 (18.5)
M. thermoautotrophicum	29 (22.8)	26 (89.7)	12 (41.4)
A. fulgidus	51 (32.3)	47 (92.2)	20 (39.2)
P. horikoshii	53 (35.3)	42 (79.3)	13 (24.5)
E. coli	113 (38.6)	102 (90.3)	30 (26.5)
H. influenzae	52 (34.0)	48 (92.3)	22 (42.3)
H. pylori	34 (28.8)	29 (85.3)	11 (32.4)
B. subtilis	107 (38.6)	101 (94.4)	32 (29.9)
M. genitalium	18 (29.5)	17 (94.4)	9 (50.0)
M. pneumoniae	25 (34.2)	23 (92.0)	14 (56.0)
M. tuberculosis	67 (31.5)	62 (92.5)	22 (32.8)
B. burgdorferi	22 (25.9)	20 (90.9)	13 (59.1)
T. pallidum	34 (35.8)	27 (79.4)	9 (26.5)
Synechocystis sp.	71 (32.6)	57 (80.3)	17 (23.9)
A. aeolicus	32 (24.6)	28 (87.5)	7 (21.9)
S. cerevisiae	101 (24.4)	60 (59.4)	36 (35.6)

[a] Percentage to the total number of ATP-binding proteins shown in Table 1.
[b] Percentage to the total number of membrane protein/ATP-binding protein pairs.

DISCUSSION

We investigated the distribution of membrane proteins in 16 complete genome sequences. We showed that statistically significant portions of membrane proteins were encoded in the genome as tandem clusters. There was a total of 1957 tandem clusters in the 16 genomes (Table 2). We analyzed the sequence similarity of membrane proteins in tandem clusters in order to identify, first, conserved (orthologous and paralogous) tandem clusters and, second, paralogous proteins within a tandem cluster. Most of the conserved clusters and/or the clusters containing paralogs represented functionally well-identified proteins (Table 3 and Fig. 2). There were eight conserved clusters whose functions were not known (Table 4). We predicted seven

Table 7. The Number of Bacteria-type ABC Transporter Components

Organism	Total number of ABC transporter components	Number of fused components[a]	Number of isolated components[b]			
			membrane component	ATP-binding component	substrate binding component	fused component
M. jannaschii	39		1	5		
M. thermoautotrophicum	36			6		
A. fulgidus	97		1	7		
P. horikoshii	85	2	1	6	3	
E. coli	242	9		6	17	4
H. influenzae	105	8	2	5	3	3
H. pylori	48	4	1	6	3	3
B. subtilis	207	12	1	13	9	4
M. genitalium	34	3		2		1
M. pneumoniae	35	3		3		1
M. tuberculosis	126	10	1	8	6	3
B. burgdorferi	44			3	2	
T. pallidum	50			7	3	
Synechocystis sp.	139	11	8	17	16	10
A. aeolicus	35	2	6	5	4	1
S. cerevisiae	38	23		7	7	22

[a] A fused component consists of membrane and ATP-binding components in most cases.
[b] A component is isolated when there is no other component within five gene positions on both sides.

transporters (Table 5) among 55 functionally unknown clusters containing paralogs, according to an empirical rule concerning the cluster size and the number of transmembrane segments. This was an attempt to use the information of structural features in functional annotation of membrane proteins without relying on sequence similarity. In addition, we identified probable membrane protein partners of isolated ATP-binding protein components in the ABC transporters by searching for adjacent pairs of membrane proteins and ATP-binding proteins.

Table 8. Predicted Membrane Protein Components of ABC Transporters

Cluster number	Organism	ATP-binding protein[a]	Membrane protein	TM
1	Af	AF0004 (+)*	AF0008 (+)	10
	Hi	HI1252 (+)*	HI1242 (−)*	10
	Bs	ExpZ (−)*	YdgK (+)	11
	Bs	YfmM (−), yfmR (+)	YfmO (+)	11
2	Mg	MG065 (+)*	MG064 (+)	9
	Mg	MG468.1 (−), MG467 (−)	MG464 (−)	6
	Mp	R02_orf465 (−)*	R02_orf1386V (−)	9
	Mp	K05_orf284 (+), K05_orf339 (+)	K05_orf385 (+), K05_orf1882 (+)	5, 7
3	Bs	EcsA (+)*	EcsB (+)	8
	Bs	YthP (−)*	YthQ (−)	8
	Ss	s110489 (−)*	slr0096 (+)	11
4	Af	AF0393 (−)*	AF0392 (−)	12
	Af	AF1136 (+), AF1139 (+)	AF1140 (+)	9
	Ph	PH1230 (+)	PH1231 (+)	11
5	Tp	TP0881 (−)*	TP0880 (−)	6
6	Af	AF1170 (+)*	AF1169 (+)	4
	Ph	PH0157 (+)	PH0159 (−)	5

[a] The strand direction shown in parentheses and isolated components in Table 7 shown with asterisks.

Our analysis depends on the accuracy of predicting membrane proteins, for which we used the TSEG program (http://www.genome.ad.jp/SIT/tseg.html). As reported previously (Kihara et al. 1998), TSEG missed 14.9% of real transmembrane segments (false negatives) and overpredicted 8.5% of nontransmembrane segments (false positives) in our test data set. Let us assume that this prediction accuracy applies to the current genome-scale analysis and that every transmembrane segment is predicted independently with the above accuracy. Then the probability that a membrane protein predicted to have three transmembrane (TM) segments that are actually a globular (nonmembrane) protein is (0.085)3 = 6.1×10^{-4} and conversely, the probability that a real 3TM protein is predicted to be a globular protein is (0.149)3 = 3.3×10^{-3}. Considering the number of predicted membrane proteins in the present analysis (Table 1), it is almost certain that those predicted to have three or more TM segments are real membrane proteins, and that real membrane proteins with three or more TM segments would not be missed. To take another example, the probability that a real 6TM protein is predicted to have three TM segments is 6C3 × (0.149)3= 0.066 and the probability that a membrane protein predicted to have nine TM segments that are indeed a 6TM protein is 9C3 × (0.085)3= 0.052. Though the probability depends on the predicted number of TM segments in a protein, these values provide an idea of how rare it is to mispredict the number of TM segments by three or more. With careful consideration of the limitation of the prediction accuracy, we think that the information of the predicted number of TM segments can be used in a positive way to understand protein functions.

The main conclusion of the present study is that about half of the membrane proteins form tandem clusters in the genome. There are several possible explanations for this observation and they are not necessarily mutually exclusive. First, the functional coupling of protein products is probably the most dominant biological constraint on such clustering in the genome. Despite the fact that the locations of orthologous genes are extensively shuffled even in the genomes of closely related species, some gene clusters are found to be tightly coupled as conserved operons (Tamames et al. 1997; Dandekar et al. 1998; Overbeek et al. 1999). The genome is viewed increasingly as a dynamic entity, and conserved gene clusters may also result from the horizontal gene transfer (Xu et al. 1998). Not including the conserved tandem clusters reported above, there may be other functionally coupled tandem clusters though they are not conserved among the organisms studied here. We expect that as more completely sequenced genomes become available the possibility of identifying functional clues will increase. Second, it is conceptually possible to imagine that the apparent clustering of membrane proteins results from the clustering of nonmembrane proteins as a background clustering. There are a number of known examples of functionally coupled nonmembrane protein clusters. However, nonmembrane proteins constitute 70%–80% of the total proteins, and the majority of nonmembrane proteins do not form gene clusters as evidenced by the extensive shuffling of orthologous genes. In fact, we believe that the functional coupling of protein products alone cannot explain the statistically significant occurrence of membrane protein gene clusters, although the functions of all membrane proteins are not yet known.

Third, we present a hypothesis that forming tandem clusters is favorable for the cellular mechanism of membrane protein expression, perhaps at the stage of protein translocation to the cell membrane. The bacterial protein translocation machinery is well studied in *E. coli*. One is the Sec machinery, which involves SecB that binds to the mature regions of nascent proteins and delivers them to SecY/E/G translocon, using the energy of ATP hydrolysis by SecA and proton motive force (Tokuda 1994). Another involves the signal recognition particle (SRP) that interacts with the hydrophobic signal peptide of a nascent protein. The two translocation pathways seem to use the common translocon (Valent et al. 1998). It has been shown that a subset of membrane proteins is dependent on the SRP pathway but others are not (Ulbrandt et al. 1997). We can speculate an implication of tandem clusters for the SecB machinery. Considering the report that SecB forms a tetramer and can bind more than one polypeptide chain (Randall et al. 1998) and also the fact that bacterial mRNA is usually polycistronic, it may be favorable for the genes of membrane proteins to be positioned tandemly, so that SecB delivers them all together like an omnibus. As for *S. cerevisiae*, we cannot reason in the same way because it does not have a SecB-like protein (Lyman and Schekman 1996), though the translocon complex is similar to that of bacteria (Jungnickel et al. 1994). Still, there may be some biological implications in *S. cerevisiae* as well, because tandem clusters of membrane proteins are as abundant as in bacteria (Table 2) without bacteria-type operon structures. Further experimental analysis of genome-scale translocation mechanism is required for the validation of our hypothesis.

Although the locations of genes and their amino acid sequences can be determined rapidly by whole genome sequencing, the functional identification of individual genes has been a slow and tedious process. We have shown that missing permease protein components of ABC transporters may be identified by searching for conserved clusters of membrane proteins and ATP-binding proteins. Generally speaking, this type of analysis extends the current knowledge on functions in terms of physical coupling of genes. Namely, when the function is known for only one of the two genes but the other gene is physically coupled, then the known function may be extended to include both genes. The analysis can be further generalized to include other types of couplings, such as identifying sequence motifs that are known to be present on two interacting proteins. Furthermore, new experimental methods in functional genomics provide direct information about coupling of genes; especially cDNA microarrays (Brown and Botstein 1999) at the level of mRNA expression and yeast two hybrid systems at the protein–protein interaction level. Thus, based on the concept of links, or binary relations (Kanehisa 2000), both computational predictions and data processing of systematic experiments can be integrated to identify functional couplings and eventually to understand the entire network of genes and proteins.

METHODS

Complete Genomes

We analyzed the complete genome sequences of the following 16 organisms: *M. jannaschii* (Bult et al. 1996), *M. thermoautotrophicum* (Smith et al. 1997), *A. fulgidus* (Klenk et al. 1997), and *P. horikoshii* (Kawarabayasi et al. 1998) from archaea, *E. coli* (Blattner et al. 1997), *H. influenzae* (Fleischmann et al. 1995), *Helicobacter pylori* (Tomb et al. 1997), *B. subtilis* (Kunst et al. 1997), *Mycoplasma genitalium* (Fraser et al. 1995), *Mycoplasma pneumoniae* (Himmelreich et al. 1996), *Mycobacterium tuberculosis* (Cole et al. 1998), *Borrelia burgdorferi* (Fraser et al. 1997), *T. pallidum* (Fraser et al. 1998), *A. aeolicus* (Deckert et al. 1998), and *Synechocystis* sp. PCC6803 (Kaneko et al. 1996) from bacteria, and *S. cerevisiae* (Goffeau et al. 1997) from eukarya. The amino acid sequence data and the information of gene locations (ORFs) were taken from the complete genomes section of GenBank (ftp://ncbi.nlm.nih.gov/genbank/genomes/) as incorporated in the GENES database in KEGG (http://www.genome.ad.jp/kegg/). We accepted the authors' ORF assignments except for *P. horikoshii*; we removed 252 shadow genes which were entirely embedded in longer ORFs on the other strand (Kawarabayasi et al. 1998). The information of functional annotations is taken from KEGG (Ogata et al. 1999) and SWISS-PROT (Bairoch and Apweiler 1998), together with some new functional assignments we made using the sequence similarity search.

Identification of Membrane Proteins

Membrane proteins were identified from the sets of ORFs by means of two complementary automatic procedures followed by manual verification (Fig. 4). One is to use discriminant function to detect highly hydrophobic regions in the amino acid sequence, whereas the other is to rely on sequence similarity to known membrane proteins. In the first procedure, an amino-terminal signal peptide has to be properly removed because it is a hydrophobic segment often mislabelled as a transmembrane segment by any predictive method of membrane proteins. The prediction of signal peptides is based on a method similar to PSORT (Nakai and Kanehisa 1991; Nakai and Horton 1999), which consists of two steps: first to identify the existence of a signal peptide by amino acid sequence features (McGeoch 1985), and then to detect the cleavage site using a weight matrix (von Heijne 1986).

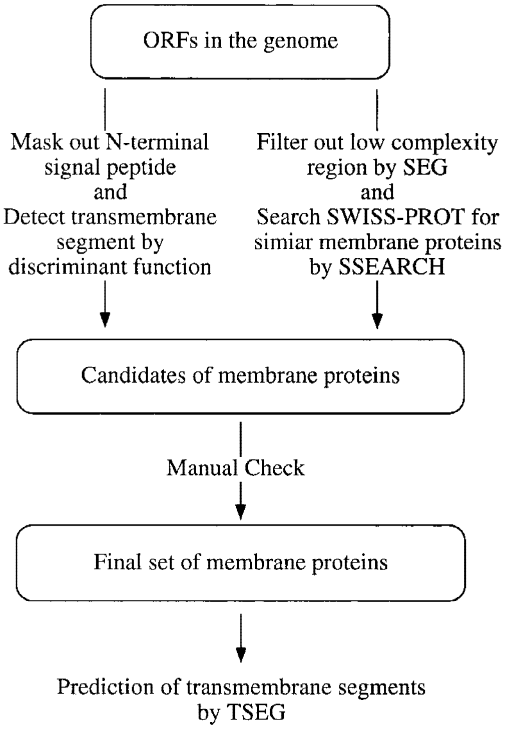
ORFs in the genome
Mask out N-terminal signal peptide and Detect transmembrane segment by discriminant function
Filter out low complexity region by SEG and Search SWISS-PROT for simiar membrane proteins by SSEARCH
Candidates of membrane proteins
Manual Check
Final set of membrane proteins
Prediction of transmembrane segments by TSEG

Figure 4 A schematic illustration of the procedures to identify membrane proteins in the complete genome.

After removal of the signal peptide, we employed discriminant analysis for distinguishing between a membrane protein and a globular (nonmembrane) protein. The discrimination function was constructed for the most hydrophobic 17-residue region of a protein sequence from the training sets of true (membrane proteins) and false (globular proteins) data (Kihara and Kanehisa 1997). The set of membrane proteins was extracted from the SWISS-PROT database release 34.0 (Bairoch and Apweiler 1998). When fragment entries were excluded and only one entry was selected from those with >30% of sequence identity, the data set contained 3251 sequences. The set of globular proteins was based on the PDBSELECT database 97-March version (Hobohm et al. 1992) excluding entries of membrane and lipid associated proteins. The 35% threshold list was used and the data set contained 928 sequences. The resulting discrimination formula was as follows:

$$f = -11.98 + 6.74 \tag{1}$$

where denotes the average hydrophobicity of the 17-residue region using the Kyte-Doolittle (1982) hydrophobicity index. An ORF is predicted to be a membrane protein if the function is positive. We chose the 17 residue-long window size because it discriminated between the two training sets best, correctly assigning 94.3% of membrane proteins and 95.2% of globular proteins (detailed data not shown).

The advantage of the first procedure is that it is an ab initio type prediction without relying on sequence similarity, but the prediction accuracy is not necessarily very high. To compensate the drawback of the first procedure, we also employed sequence similarity searches in the second procedure (Fig. 4). We used the SSEARCH program (Pearson 1991) against the SWISS-PROT database with the default parameter setting, after preprocessing of the query sequence using the SEG program (Wooton and Federhen 1993) with the default setting. SEG is effective to filter out low-complexity regions, which are stretches of hydrophobic amino acids in our case, and to reduce spurious hits (Bork and Koonin 1998). Each ORF is used as a query sequence, and it is considered as a possible membrane protein when the database hit includes any of the definite membrane proteins satisfying either of the following two criteria. The first criterion is that the E value does not exceed 0.001 (Brenner et al. 1995). In the other criterion, the E value may be ⩽0.1, but the Smith-Waterman score must be ⩾120, the Z-score must be ⩾120, and the alignment overlaps ⩾50% of the sequence lengths. The ORF sequences selected by the two procedures were then checked manually.

Prediction of Transmembrane Segments

We used the TSEG program (Kihara et al. 1998) for prediction of transmembrane segments. The basic idea of the method is to classify transmembrane segments into five types according to the average hydrophobicity and the periodicity of hydrophobicity and to model the membrane protein group with a specific number of transmembrane segments in terms of a series of different transmembrane segment types. For example, the models of membrane protein groups with one to nine transmembrane segments are, respectively, (1), (2,2), (2,3,3), (2,2,3,2), (2,3,3,2,2), (2,2,2,2,3,2), (2,3,3,2,2,2,5), (2,3,2,2,2,2,4,2), and (2,2,2,2,2,2,4,5,2) where 1–5 represent the types of transmembrane segments, 1 being most hydrophobic and 5 being least hydrophobic. The prediction of transmembrane segments involves selection of the most compatible model among different models, including one for globular proteins.

Identification of ATP-binding Proteins

The identification of ATP-binding proteins was based on the motif search. If an ORF sequence contains the P-loop ATP/GTP binding motif of the PROSITE database (Hofmann et al. 1999):

[AG]-x(4)-G-K-[ST]

then it is considered as an ATP-binding protein.

Identification of Conserved Clusters

SEG and SSEARCH were used to identify membrane proteins and ATP-binding proteins that correspond to each other in two clusters. Two clusters were considered to be conserved if >1 pair of constituent proteins is similar. The Smith-Waterman score of 150 was used for the similarity criterion. Generally, the threshold score of 120 is high enough to detect related sequences (Pearson 1996, 1998). However, according to our experience, obviously unrelated sequences were found for both membrane and ATP-binding protein searches when the threshold score of 120 was used. Conserved clusters were collected into a group by single-linkage clustering; namely, a cluster was added to the group if it was similar to at least one of the clusters in the group.

ACKNOWLEDGMENTS

This work was supported in part by a Grant-in-Aid for Scientific Research on the Priority Area Genome Science from the Ministry of Education, Science, Sports, and Culture of Japan. The computational resource was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Arkin, I.T., A.T. Brünger, and D.M. Engelman. 1997. Are there dominant membrane protein families with a given number of helices? *Proteins: Struct. Funct. Genet.* **28:** 465–466.

Aurora, R. and G.D. Rose. 1998. Seeking an ancient enzyme in *Methanococcus jannaschii* using ORF, a program based on predicted secondary structure comparisons. *Proc. Natl. Acad. Sci.* **95:** 2818–2823.

Bairoch, A. and R. Apweiler. 1998. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res.* **26:** 38–42.

Blattner, F.R., G. Plunkett, III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1462.

Bork, P. and E.V. Koonin. 1998. Predicting functions from protein sequences– where are the bottlenecks? *Nature Genet.* **18:** 313–318.

Boyd, D., C. Schierle, and J. Beckwith. 1998. How many membrane proteins are there? *Protein Sci.* **7:** 201–205.

Brenner, S.E., T. Hubbard, A. Murzin, and C. Chothia. 1995. Gene duplications in *H. influenzae*. *Nature* **378:** 140.

Brown, P.O. and D. Botstein. 1999. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* **21:** 33–37.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. Fitzgerald, R.A. Clayton, J.D. Gocayne et al. 1996. Complete genome sequence of the Methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058–1073.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry, III et al. 1998. Deciphering the biology of *Mycrobacterium tuberculosis* from the complete genome sequence. *Nature* **393:** 537–544.

Dandekar, T., B. Snel, M. Huynen M, and P. Bork. 1998. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23:** 324–328.

Dean, M. and R. Allikmets. 1995. Evolution of ATP-binding cassette transporter genes. *Curr. Opin. Genet. & Development.* **5:** 779–785.

Deckert, G., P.V. Warren, T. Gaasterland, W.G. Young, A.L. Lenox, D.E. Graham, R. Overbeek, M.A. Snead, M. Keller, M. Aujay et al. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392:** 353–358.

Fath, M.J. and R. Kolter. 1993. ABC transporters: Bacterial exporters. *Micobiol. Rev.* **57:** 995–1017.

Fetrow, J.S. and J. Skolnick. 1998. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T_1 ribonucleases. *J. Mol. Biol.* **281:** 949–968.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.-F. Tomb, B.A. Dougherty, J.M. Merrick et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science* **269:** 496–512.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397–403.

Fraser, C.M., S. Casjens, W.M. Huang, G.G. Sutton, R. Clayton, R. Lathigra, O. White, K.A. Ketchum, R. Dodson, E.K. Hickey et al. 1997. Genomic sequence of a lyme disease spirochete, *Borrelia burgdorferi*. *Nature* **390:** 580–586.

Fraser. C,M,, S.J. Norris, G.M. Weinstock, O. White, G.G. Sutton, R. Dodson, M. Gwinn, E.K. Hickey, R. Clayton, K.A. Ketchum et al. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281:** 375–388.

Frishman, D. and H.W. Mewes. 1997. Protein structural classes in five complete genomes. *Nature Struct. Biol.* **4:** 626–628.

Froshauer, S., G.N. Green, D. Boyd, K. McGovern, and J. Beckwidth. 1988. Genetic analysis of the membrane insertion and topology of MalF, a cytoplasmic membrane protein of *Escherichia coli. Mol. Biol.* **200:** 501–511.

Goffeau A., R. Aert, M.L. Agostini-Carbone, A. Ahmed, M. Aigle, L. Alberghina, K. Albermann, M. Albers, M. Aldea, D. Alexandraki et al. 1997. The yeast genome directory. *Nature* **387**(6632 Suppl.).

Higgins, C.F. 1992. ABC transporters. From microorganisms to man. *Annu. Rev. Cell. Biol.* **8:** 67–113.

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B.-C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acid Res.* **24:** 4420–4449.

Hobohm, U., M. Scharf, R. Schneider, and C. Sander. 1992. Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Prot. Sci.* **1:** 409–417.

Hofmann, K., P. Bucher, L. Falquet, and A. Bairoch. 1999. The PROSITE database, its status in 1999. *Nucleic Acid Res.* **1:** 215–219.

Jones D.T. 1998. Do transmembrane protein superfold exist? *FEBS Lett.* **423:** 281–285.

Jungnickel, B., T.A. Rapoport, E. Hartmann. 1994. Protein translocation: common themes from bacteria to man. *FEBS Lett.* **346:** 73–77.

Kanehisa, M. 2000. Pathway databases and higher order function. *Adv. Protein Chem.* **54:** in press.

Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto et al. 1996. Sequence analysis of the genome of the unicellular Cyanobacterium *Synechocystis* sp. Strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3:** 109–136.

Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama et al. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5:** 55–76.

Kihara, D. and M. Kanehisa. 1997. Detection of membrane proteins in the whole genome sequences. *Genome Informatics 1997*. pp.300–301.Universal Academy Press, Tokyo, Japan.

Kihara, D., T. Shimizu, and M. Kanehisa. 1998. Prediction of membrane proteins based on classification of transmembrane segments. *Protein Eng.* **11:** 961–970.

Klenk, H.P., R.A. Clayton, J.-F. Tomb, O. White, K.E. Nelson, K.A. Ketchum, R.J. Dodson, M. Gwinn, E.K. Hickey, J.D. Peterson et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390:** 364–370.

Kunst, F., N. Ogasawara, I. Moszer, A.M. Albertini, G. Alloni, V. Azevedo, M.G. Bertero, P. Bessieres, A. Bologin, S. Borchert et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390:** 249–256.

Kyte, J. and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157:** 105–132.

Lyman, S.K. and R. Schekman. 1996. Polypeptide translocation machinery of the yeast endoplasmic reticulum. *Experientia* **52:** 1042–1049.

McGeoch, D.J. 1985. On the predictive recognition of signal peptide sequences. *Virus Res.* **3:** 271–286.

Nakai, K. and M. Kanehisa. 1991. Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins: Struct. Funct. Genet.* **11:** 95–110.

Nakai, K. and P. Horton. 1999. PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **24:** 34–6.

Overbeek, R., M. Fonstein, M. D'Souza, G.D. Pusch, and N. Maltsev. 1999. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96:** 2896–2901.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acid Res.* **27:** 29–34.

Paulsen, I.T., M.K. Sliwinski, and H. Saier, Jr. 1998. Microbial genome analyses: Global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol.* **277:** 573–592.

Pearson, W.R. 1991. Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11:** 635–650.

———1996. Effective protein sequence comparison. *Methods Enzymol.* **266:** 227–258.

———1998. Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276:** 71–84.

Randall, L.L., S.J.S. Hardy, T.B. Topping, V.F. Smith, J.E. Bruce, and R.D. Smith. 1998. The interaction between the chaperone SecB and its ligands: Evidence for multiple subsites for binding. *Protein Sci.* **7:** 2384–2390.

Saurin, W. and E. Dassa. 1994. Sequence relationships between integral inner membrane proteins of binding protein-dependent transport systems: Evolution by recurrent gene duplications. *Protein Sci.* **3:** 325–344.

Smith, D.R, L.A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert et al. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: Functional analysis and comparative genomics. *J. Bacteriol.* **179:** 7135–7155.

Tam, R. and M.H. Saier, Jr. 1993. Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. *Microbiol. Rev.* **57:** 320–346.

Tamames, J., G. Casari, C. Ouzounis, and A. Valencia. 1997. Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44:** 66–73.

Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.

Tokuda, H. 1994. Biochemical characterization of the presecretory protein translocation machinery of *Escherichia coli*. *FEBS Lett.* **346:** 65–68.

Tomb, J.-F., O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty et al. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388:** 539–547.

Tomii, K. and M. Kanehisa. 1998. A comparative analysis of ABC transporters in complete microbial genomes. *Genome Res.* **8:** 1048–1059.

Ulbrandt, N.D., J.A. Newitt, and H.D. Bernstein. 1997. The *E. coli* signal recognition particle is required for the insertion of a subset of inner membrane proteins. *Cell* **88:** 187–196.

Valent, Q.A., P.A. Scotti, S. High, J.W. de Gier, G. von Heijne, G. Lentzen, W. Wintermeyer, B. Oudega, and J. Luirink. 1998. The *Escherichia coli* SRP and SecB targeting pathways converge at the translocon. *EMBO J.* **17:** 2504–2512.

von Heijne, G. 1986. A new method for predicting signal sequence cleavage sites. *Nucleic Acid Res.* **14:** 4683–4690.

Wallin, E. and G. von Heijne. 1998. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7:** 1029–1038.

Wooton, J.C. and S. Federhen. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Computers and Chem.* **17:** 149–163.

Xu, Y., B.E. Murray, and G.M. Weinstock. 1998. A cluster of genes involved in polysaccharide biosynthesis from *Enterococcus faecalis* OG1RF. *Infect. Immun.* **66:** 4313–4323.

Zhang, X. and T.F. Smith. 1998. Yeast "Operons". *Microbial Comp. Gen.* **3:** 133–140.

Received September 27, 1999; accepted in revised form March 27, 2000.