J Mol Evol (1998) 47:200–210

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 1998

Molecular Evolution of the Domain Structures of Protein Disulfide Isomerases

Satoru Kanai,[1] Hiroyuki Toh,[1] Toshiya Hayano,[2] Masakazu Kikuchi[3]

[1] Department of Bioinfomatics, Biomolecular Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka, 565 Japan
[2] Environment Research Group, Fundamental Research Laboratories, Corporate Research and Development Laboratory, Tonen Corporation, 1-3-1 Nishi-tsurugaoka, Ohi-machi, Iruma-gun, Saitama, 365 Japan
[3] Department of Bioscience and Technology, Faculty of Science and Engineering, Ristumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, 525-77 Japan

Abstract. Protein disulfide isomerase (PDI) is an enzyme that promotes protein folding by catalyzing disulfide bridge isomerization. PDI and its relatives form a diverse protein family whose members are characterized by thioredoxin-like (TX) domains in the primary structures. The family was classified into four classes by the number and the relative positions of the TX domains. To investigate the evolution of the domain structures, we aligned the amino acid sequences of the TX domains, and the molecular phylogeny was examined by the NJ and ML methods. We found that all of the current members of the PDI family have evolved from an ancestral enzyme, which has two TX domains in the primary structure. The diverse domain structures of the members have been generated through domain duplications and deletions.

Key words: Evolution — Protein — Protein disulfide isomerase — Thioredoxin

Introduction

The information for the tertiary structure of a protein is basically contained in its amino acid sequence. A string of amino acid residues is folded into a unique tertiary structure according to the program encoded by the sequence (Anfinsen 1973). However, several protein factors involved in protein folding have been identified recently. These factors do not change the directions for protein folding encoded by the amino acid sequences but assist in the structure formation by inhibiting incorrect folding or promoting correct folding (reviewed by Gething and Sambrook 1992).

Protein disulfide isomerase (PDI) is one such factor, which was first identified as an enzyme that promoted protein folding by catalyzing the isomerization of disulfide bonds (Freedman 1989; Freedman et al. 1989). But, PDI has also been found to have chaperone-like activity. It prevents intermediates in the protein folding pathway from aggregation and assists their correct folding (Wang and Tsou 1993; Cai et al. 1994; Puig and Gilbert 1994). Thus, PDI and its relatives are involved in protein folding in two different ways.

PDI and its relatives constitute a protein family (see review, Freedman et al. 1994). Cloning and sequencing analyses have revealed the primary structures of diverse members of the family. One of the characteristics of this protein family is that the members have two or three TX domains in their primary structures. Each domain includes the active site for disulfide bridge isomerization (Rupp et al. 1994).

We have classified PDIs and their relatives into four groups, based on the number and the relative positions of the TX domains (Fig. 1). Class 1 is the major constituent

Abbreviations: PDI, protein disulfide isomerase; TX, thioredoxin-like; NJ, neighbor-joining; ML, maximum likelihood; AIC, Akaike information criterion

Correspondence to: S. Kanai; *e-mail:* skanai@beri.co.jp

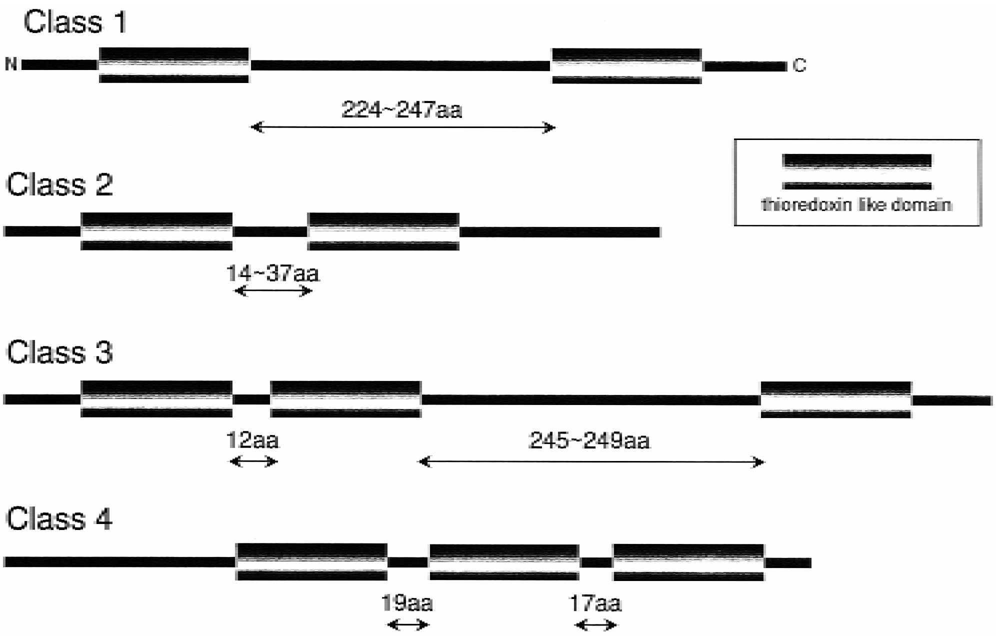

Fig. 1. Domain structures of PDIs.

of the family, whose members have TXs at both the N- and C-terminal regions. The sizes of the members are about 500 amino acid residues in length, and the two TX domains are connected by a polypeptide of about 200 amino acid residues. The members of the family are found in animals, plants, and fungi. Thirty amino acid sequences of this class are available now. The members of class 2 are about 400 amino acid residues in length and have two TX domains in their primary structures. However, the relative positions of the domains of the class 2 proteins are different from those of the class 1 proteins. Both TX domains are present at almost the middle of the primary structure and are connected by a 30-amino-acid polypeptide. Six-amino-acid sequences of this class have been determined. The proteins belonging to class 3 have three TX domains. The first and second TX domains are tandemly repeated at the N-terminal region, with a connecting polypeptide of about 10 residues, while the third domain is present at the C-terminal region. The polypeptide connecting the second and the third TX domains is about 200 residues in length. Four amino acid sequences of this class are available now. Class 4 has only one member, which also has three TX domains. However, the distribution of the three domains is different from those of the class 3 proteins. The three domains are tandemly placed and are connected by polypeptides of about 20 residues.

The evolutionary relationships among the members of the protein family were outlined in a review article by Freedman et al. (1994). Sahrawy et al. (1996) investigated the positions of the introns in thioredoxins and thioredoxin-like domains. In their paper, they classify the PDI family into two groups: members with two TX domains and those with three TX domains. They also showed a phylogenetic tree of TX domains and thioredoxins. These analyses provide an overview of the molecular evolution of the protein family. However, neither group performed a statistical evaluation of the obtained evolutionary relationships of the family. In this paper, we focus on the molecular evolution of the domain structures of the family and evaluate the statistical reliability of evolutionary relationships thus obtained. We found that the domain structures of PDI relatives have been independently reorganized on many occasions.

Materials and Methods

Sequence Data. There are many sequences of PDIs and their relatives available for molecular phylogenetic analysis. However, it is difficult to include all of the data for the current analysis, mainly due to the large amount of computational time required for the construction and the statistical evaluation of the molecular phylogeny. Therefore, we neglected the closely related members and selected a small number of representatives of the four classes. As described above, class 1 is the major constituent of the family, and is further divided into four subclasses. We selected four amino acid sequences from the subclasses as the representatives: human PDI, yeast PDI, trypanosoma BS2, and human erp60. Class 2 is composed of highly diverse members. We selected four amino acid sequences as the representatives: erp5 and its relatives from humans, *Caenorhabditis elegans,* amoeba, and alfalfa. Class 3 consists of the mammalian erp72s and the *C. elegans* counterpart. We used the human and *C. elegans* erp72 amino acid sequences. Class 4 includes only one sequence, human PDIR. Information about the data, including references and sources, is listed in Table 1.

Multiple Sequence Alignment. A multiple alignment was constructed with the program Clustal W (Higgins et al. 1991; Thompson et al. 1994). The obtained alignment was modified a little by visual inspection to accommodate the gap positions by considering the secondary structures. An alignment editor, Seaview (Galtier et al. 1996), was

Table 1. List of the proteins used in the analyses[a]

		Organism		N-terminal domain				C-terminal domain			
Class	Seq name	*Formal name* (Common name)	Reference	AA	Name	S	L	Name	S	L	Dnc
1	BS2_TRYBB	*Trypanosoma brucei brucei*	Hsu et al. 1989	483	bs2trybbN	8	101	bs2trybbC	338	102	229
	ER60_HUMAN	*Homo sapiens*	Bourdi et al. 1995	481	er60humanN	4	104	er60humanC	355	105	247
	PDI_HUMAN	*Homo sapiens*	Tasanen et al. 1988	489	pdihumanN	8	106	pdihumanC	351	104	237
	PDI_YEAST	*Saccharomyces cerevisiae*	Scherens et al. 1992	494	pdiyeastN	7	104	pdiyeastC	351	105	240
2	2024291A	*Acanthamoeba castellanii*	Wong and Bateman 1994	406	2024291aN	32	100	2024291aC	165	100	33
	ERP5_CAEEL	*Caenorhabditis elegans*	Wilson et al. 1994	440	erp5caee1N	27	103	erp5caee1C	167	105	37
	ERP5_HUMAN	*Homo sapiens*	Hayano and Kikuchi 1995a	440	erp5humanN	28	103	erp5humanC	163	106	32
	ERP5_MEDSA	*Medicago sativa* (alfalfa)	Shorrosh and Dixon 1992	336	erp5medsaN	4	104	erp5medsaC	122	105	14

					1st domain			2nd domain			3rd domain				
					Name	S	L	Name	S	L	Name	S	L	D12	D23
3	ER72_CAEEL	*Caenorhabditis elegans*	Wilson et al. 1994	664	er72caeel1	85	99	er72caeel2	196	103	er72caeel3	548	106	12	249
	ER72_HUMAN	*Homo sapiens*	Huang et al. 1991	625	er72human1	45	103	er72human2	160	103	er72human3	508	107	12	245
4	PDIR_HUMAN	*Homo sapiens*	Hayano and Kikuchi 1995b	519	pdirhuman1	155	105	pdirhuman2	279	104	pdirhuman3	400	105	19	17

[a] ‘‘AA’’ denotes the length of the protein (amino acid residues). ‘‘S’’ indicates the site number at which the TX domain starts. ‘‘L’’ indicates the length of the TX domain. ‘‘Dnc’’, ‘‘D12’’, and ‘‘D23’’ indicate the number of amino acid residues between two adjacent TX domains, respectively

used for the modification. Finally, the alignment sites with the gaps were removed from the alignment for the molecular phylogenetic studies.

Molecular Phylogeny. To construct the molecular phylogeny, the NJ method (Saitou and Nei 1987) and the ML method (Felsenstein 1981; Kishino et al. 1990) were used. The bootstrap analysis (Felsenstein 1985) was done with 1,000 iterations of resamplings and tree reconstructions. The PAM001 (Dayhoff et al. 1978) was used to calculate the genetic distance for the NJ analysis. On the other hand, the Dayhoff model (Dayhoff et al. 1978) was adopted as the evolutionary model for the ML analysis, because most of the trees showing minimal AIC (Akaike 1974) suggested the Dayhoff model in the preliminary analyses. The molecular phylogeny studies were done with the program packages PHYLIP (Felsenstein 1993, 1996) and MOLPHY (Adachi and Hasegawa 1996). The trees were drawn by TreeTool (Maciukenas and McCaughey 1994).

Results and Discussion

Class 1 Has the Primary Domain Structure of PDI

Figure 2 shows an NJ tree of the TX domains. We tried to determine the root position of the tree by including the amino acid sequences of thioredoxins as the outgroup. Unfortunately, the root thus obtained did not show statistical significance, probably due to the high sequence divergence and the small number of alignment sites. Therefore, we do not show the root of the tree. On the other hand, we found, in the above approach, that the thioredoxins formed a single cluster, which was statistically distinct from the cluster of the TX domains of the PDI family (data not shown). Our observation suggested that the members of the PDI family did not appear independently from the fusion of thioredoxins, but that all of them are derived from a common ancestral PDI.

Our tree with the TX domains and the thioredoxins is similar to that previously constructed by Sahrawy et al. (1996). However, their tree did not fully cover the variety of the PDI family. For example, their tree included only one class 2 sequence, in spite of the diversity of class 2. In addition, class 4 was not considered in their analysis.

We classified the members of the PDI family into four groups on the basis of the morphology of the domain structures (see Fig. 1). As described above, all of these diverse structures are considered to be derived from a common ancestral PDI. One of our interests was to determine which class has the most primary domain structure of PDI. In other words, we would like to identify the domain structure of the most ancestral PDI. We investi-

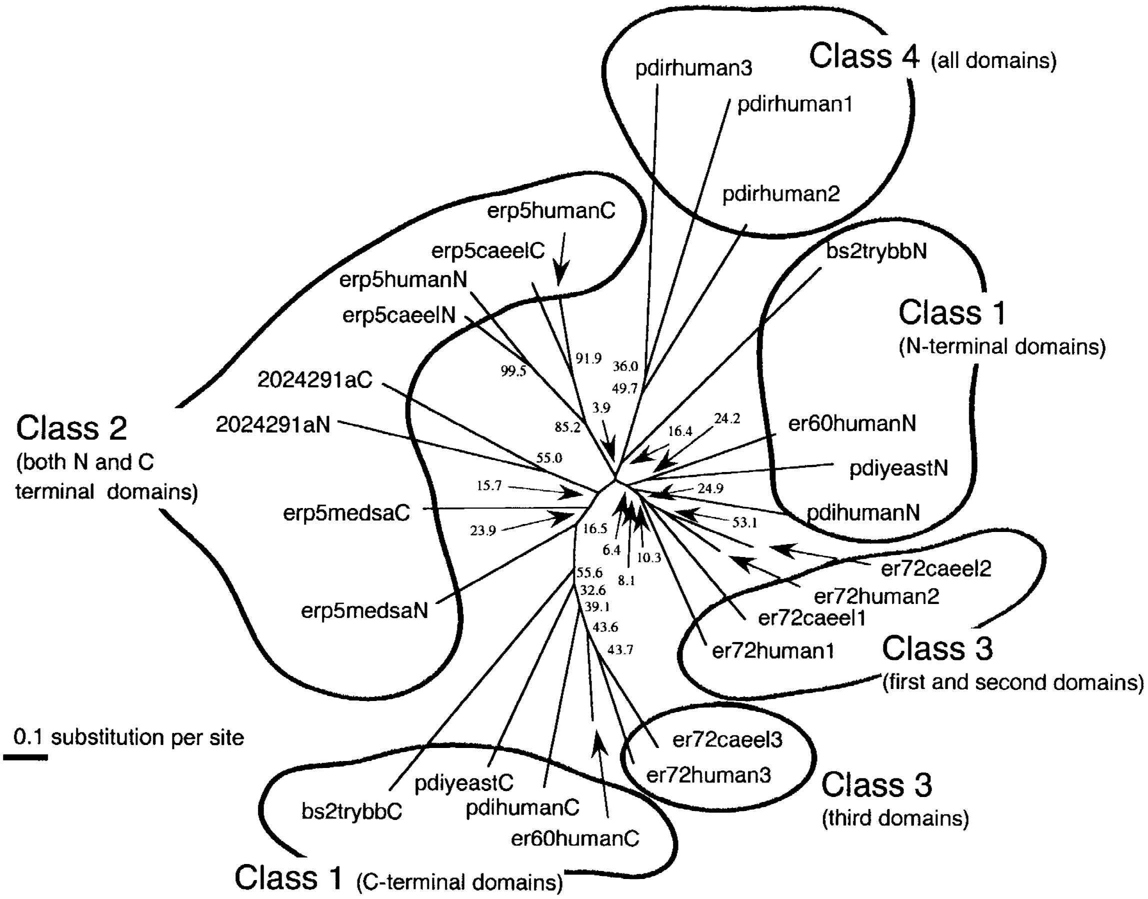

Fig. 2. An unrooted phylogenetic tree obtained by the neighbor-joining method. The *numbers* at the nodes indicate the bootstrap probabilities. The rules of name abbreviation are listed in Table 1.

gated this point with the NJ tree shown in Fig. 2. Generally speaking, a classification by a morphological viewpoint does not always correspond with that by molecular evolutionary viewpoint. However, the clusters found in the tree roughly corresponded with the morphological classification, except for class 2. The N- and C-terminal TX domains of class 1 each formed single clusters in the tree, respectively. To clarify the statistical significance of the clustering, we aligned only the N- and C-terminal TX domains of class 1. The NJ and ML trees both consisted of two clusters, corresponding to the N- and C-terminal domains. The bootstrap probability for this clustering by the NJ method was 94.1% and that by the ML method was 99.0%. Thus, the clustering pattern of class 1 is regarded as being statistically significant. On the other hand, the topology in the cluster of the N-terminal domains was slightly different from that of the C-terminal domains. The difference might be attributed to the high sequence divergence and the small number of alignment sites. In this paper, we will focus on the evolution of the domain structures, and the evolutionary relationships within each class will not be discussed. The cluster of the first and second domains of class 3 was included in the cluster of the class 1 N-terminal domains. Similarly, the third domain of class 3 was present in the cluster of the class 1 C-terminal domains. The three TX domains of class 4 formed a single cluster, which was found in the cluster of the class 1 N-terminal domain. The observation suggests that class 3 and class 4 are derived from class 1. In contrast, class 2 did not form a single cluster, which suggests that the class is a mixture of PDIs with different evolutionary origins. The figure shows that class 2 is divided into three subclasses. One of them includes the N- and C-terminal domains of the human and *C. elegans* erp5s. The N- and C-terminal domains of the amoeba erp5 homologue form the second subclass. The N- and C-terminal domains of the alfalfa erp5, the third subclass of class 2, are also closely related, but do not form a cluster in the figure. Thus, the members of class 2 appeared from three independent origins, although the N- and C-terminal domains of class 2 are more closely related to each other than to those of class 1. In other words, the divergence of the N- and C-terminal domains of class 1 is more ancient than that of any subclass of class 2. Due to the failure of the root assignment, the evolutionary relationship between class 1 and class 2 is ambiguous in this tree. However, more detailed analyses by the ML method, which will be described below, suggested that both the N- and C-terminal TX domains of class 2 belong to the cluster of class 1 N-terminal TX domains.

Sahrawy et al. (1996) classified the PDI family into two groups based on the domain structures, the members with two TX domains and those with three TX domains. The two groups roughly correspond with class 1 and class 3 in our classification. They determined the root of the TX domains, using thioredoxins as the outgroup, which divided the N- and C-terminal TX domains of class 1. Like our tree, the TX domains of class 3 are included in those of class 1. However, they did not evaluate the statistical significance of the tree thus obtained.

Our observations, together with the previous tree determined by Sahrawy et al. (1996), suggest that the domain structure of the most ancestral PDI corresponds with that of class 1 and that the domain structures of the other classes are derived from the class-1-type structure. The next question was how the other domain structures evolved from the class-1-like structure. The bootstrap probabilities of the nodes in Fig. 1 were not always high. Therefore, the evolutionary scenario described above should be confirmed by another approach. We examined the evolutionary positions of the other three classes by the ML method.

Evolutionary Positions of the Class 2 TX Domains

The tree shown in Fig. 2 revealed three evolutionary problems with the domain structures of class 2. Class 2 was divided into three subgroups, which seemed to have appeared independently. So, the first question was whether class 2 is an artificial classification without any evolutionary meaning. The N- and C-terminal TX domains of each class 2 subgroup seem to be more closely related to each other than to those of class 1. In other words, the divergence between the N- and C-terminal domains of class 2 seems to have occurred relatively late, as compared to the early divergence between the N- and C-terminal domains of class 1. The second question was whether the difference in the divergence pattern of the TX domains between class 1 and class 2 is statistically significant. The third question was how the domain structure of class 2 evolved from the class-1-like structure. It was difficult to make any definite statement about these problems based on the NJ analysis shown in Fig. 2, due to the low bootstrap probabilities of the nodes related to these problems. To examine these problems by a different approach, we tried a series of quartet tests using the ML method. Table 2 summarizes the results of the quartet tests. The three problems are interrelated. Therefore, we will explain the obtained results at first, instead of answering each question one by one. After that, we will propose a model for the evolution of class 2 to answer the questions.

In categories 1 through 8 of Table 2, the first component of each quartet was an N- or C-terminal domain of a class 2 protein. The second and third components of the quartet were the N- and C-terminal domains of a class 1 protein, respectively. The fourth component was the C-terminal domain of the other class 1 protein, which was selected to be most distantly related to the third component in Fig. 2. All of the results, except for that in category 8, suggested the topology ((1, 2), (3, 4)) as the ML tree, which means that the N- and C-terminal domains are derived from the N-terminal domains of the class-1-like structure.

In categories 9 through 12 of Table 2, the first and second components of each quartet were the N- and C-terminal domains of a class 2 protein, while the third and fourth components corresponded with the N- and C-terminal domains of a class 1 protein, respectively. Most of the results shown in categories 9 through 11 suggested the topology ((1, 2), (3, 4)) as the ML tree—that is, the divergence between the two domains of class 2 occurred independently from that of class 1. On the other hand, the results shown in category 12 suggested that the divergence pattern of the TX domains of alfalfa was different from those of the other class 2 proteins but was identical to those of the class 1 proteins.

Categories 13 through 15 of Table 2 show the results for the checks of independence in the domain duplication within class 2. The first and second components of each quartet were the N- and C-terminal TX domains of a class 2 protein, while the third and fourth components were the N- and C-terminal domains of another class 2 protein. The results shown in category 13 suggested that the domain duplications of the amoeba erp5 homologue are independent from those of the other class 2 proteins. In contrast, the first quartet test in category 14 indicated that the human and *C. elegans* erp5 proteins are derived from a common ancestral class 2 protein that is different from the ancestors of the amoeba and alfalfa erp5 proteins.

The results of the quartet tests suggested that class 2 is an artificial classification, composed of three subgroups evolved from different origins. One of them is the alfalfa erp5. Figure 2 shows the early divergence of the protein from an ancestral class 1. However, we were not able to construct a model for the domain evolution of alfalfa erp5 that could consistently explain the results of categories 7, 8, and 12 of Table 2. Therefore, we will not discuss the protein further. The second subgroup included the erp5 homologue from an amoeba. Figure 3 shows a possible evolutionary scenario of the protein. At first, the N-terminal domain of an ancestral class 1 protein was duplicated, and the protein obtained three TX domains. Then, the C-terminal domain was deleted, and the two-domain structure was reinstated. This model can explain both the relatively late divergence of the domains

Table 2. Results of quartet tests for class 2 PDIs[a]

					Topology					
	Sequence				((1,2),(3,4))		((1,3),(2,4))		((1,4),(2,3))	
Category	1	2	3	4	diff AIC	Boot P	diff AIC	Boot P	diff AIC	Boot P
1	2024291aN	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.8970	11.6	0.0047	10.5	0.0983
	↓	pdihumanN	pdihumanC	bs2trybbC	0.0	0.8159	5.5	0.1635	6.2	0.0206
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.6517	4.0	0.3450	9.1	0.0033
	↓	er60humanN	er60humanC	↓	0.0	0.9022	11.9	0.0846	12.5	0.0132
2	2024291aC	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.8721	8.6	0.0705	8.6	0.0574
	↓	pdihumanN	pdihumanC	bs2trybbC	0.0	0.7818	8.1	0.0638	7.2	0.1544
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.8834	13.5	0.1094	17.0	0.0072
	↓	er60humanN	er60humanC	↓	0.0	0.8978	16.9	0.0018	14.8	0.1004
3	erp5caeelN	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.8753	6.5	0.1012	6.7	0.0235
	↓	pdihumanN	pdihumanC	bs2trybbC	0.0	0.9631	11.7	0.0160	11.7	0.0209
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.9196	8.6	0.0444	8.6	0.0360
	↓	er60humanN	er60humanC	↓	0.0	0.9747	22.1	0.0152	22.1	0.0101
4	erp5caeelC	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.9265	7.9	0.0482	7.9	0.0253
	↓	pdihumanN	pdihumanC	bs2trybbC	0.0	0.7846	5.9	0.2071	7.8	0.0083
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.9815	20.9	0.0165	21.2	0.0020
	↓	er60humanN	er60humanC	↓	0.0	0.9640	19.9	0.0040	19.9	0.0320
5	erp5humanN	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.9804	19.2	0.0102	19.2	0.0094
	↓	pdihumanN	pdihumanC	bs2trybbC	0.0	0.9477	10.2	0.0339	10.2	0.0184
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.9465	13.5	0.0490	13.8	0.0045
	↓	er60humanN	er60humanC	↓	0.0	0.9673	23.0	0.0290	23.2	0.0037
6	erp5humanC	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.9729	15.1	0.0234	15.1	0.0037
	↓	pdihumanN	pdihumanC	bs2trybbC	0.0	0.9619	12.1	0.0226	12.1	0.0155
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.9750	16.3	0.0107	16.3	0.0143
	↓	er60humanN	er60humanC	↓	0.0	0.9848	29.4	0.0006	29.0	0.0146
7	erp5medsaN	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.7472	5.3	0.2441	7.5	0.0087
	↓	pdihumanN	pdihumanC	bs2trybbC	0.0	0.7506	5.7	0.0645	4.8	0.1849
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.8148	5.1	0.3333	4.8	0.1519
	↓	er60humanN	er60humanC	↓	0.0	0.8989	12.2	0.0606	12.5	0.0405
8	erp5medsaC	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.8201	6.3	0.0826	6.2	0.0973
	↓	pdihumanN	pdihumanC	bs2trybbC	2.5	0.1930	0.0	0.7059	2.8	0.1011
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.7729	4.4	0.2097	5.2	0.0174
	↓	er60humanN	er60humanC	↓	0.0	0.8506	8.5	0.1186	8.8	0.0308
9	2024291aN	2024291aC	bs2trybbN	bs2trybbC	0.9	0.4463	0.0	0.5230	4.1	0.0307
	↓	↓	pdihumanN	pdihumanC	2.5	0.3219	4.5	0.0418	0.0	0.6363
	↓	↓	pdiyeastN	pdiyeastC	0.0	0.8933	18.9	0.0030	15.6	0.1037
	↓	↓	er60humanN	er60humanC	0.0	0.8592	13.3	0.0194	11.5	0.1214
10	erp5caeelN	erp5caeelC	bs2trybbN	bs2trybbC	0.0	0.6689	3.0	0.0449	2.1	0.2862
	↓	↓	pdihumanN	pdihumanC	0.0	0.6580	7.0	0.0860	4.7	0.2560
	↓	↓	pdiyeastN	pdiyeastC	0.0	0.9211	12.3	0.0124	11.9	0.0665
	↓	↓	er60humanN	er60humanC	1.8	0.4420	0.0	0.5484	9.6	0.0096
11	erp5humanN	erp5humanC	bs2trybbN	bs2trybbC	0.0	0.6563	2.7	0.3241	4.8	0.0196
	↓	↓	pdihumanN	pdihumanC	0.0	0.9607	16.9	0.0076	16.9	0.0317
	↓	↓	pdiyeastN	pdiyeastC	0.0	0.9493	19.1	0.0030	18.3	0.0477
	↓	↓	er60humanN	er60humanC	0.0	0.9333	14.8	0.0604	15.5	0.0063
12	erp5medsaN	erp5medsaC	bs2trybbN	bs2trybbC	4.5	0.0249	3.4	0.2708	0.0	0.7043
	↓	↓	pdihumanN	pdihumanC	7.1	0.0200	0.0	0.8108	6.4	0.1692
	↓	↓	pdiyeastN	pdiyeastC	2.8	0.0937	0.0	0.5538	1.4	0.3525
	↓	↓	er60humanN	er60humanC	17.1	0.0419	0.0	0.9476	17.6	0.0105
13	2024291aN	2024291aC	erp5caeelN	erp5caeelC	0.0	0.9871	27.3	0.0106	27.3	0.0023
	↓	↓	erp5humanN	erp5humanC	0.0	0.9960	33.8	0.0018	33.8	0.0022
	↓	↓	erp5medsaN	erp5medsaC	0.0	0.9765	22.8	0.0190	22.8	0.0045
14	erp5caeelN	erp5caeelC	erp5humanN	erp5humanC	42.6	0.0024	0.0	0.9976	42.8	0.0000
	↓	↓	erp5medsaN	erp5medsaC	0.0	0.9201	20.6	0.0793	24.2	0.0006
15	erp5humanN	erp5humanC	erp5medsaN	erp5medsaC	0.0	0.9900	29.6	0.0075	29.6	0.0025

[a] The ‘‘diff AIC’’ denotes the difference between the minimal AIC and the AIC of each quartet. Therefore, the ‘‘diff AIC’’ is 0.0 when the quartet is the topology with the minimal AIC. ‘‘Boot P’’ indicates the bootstrap probability of each quartet

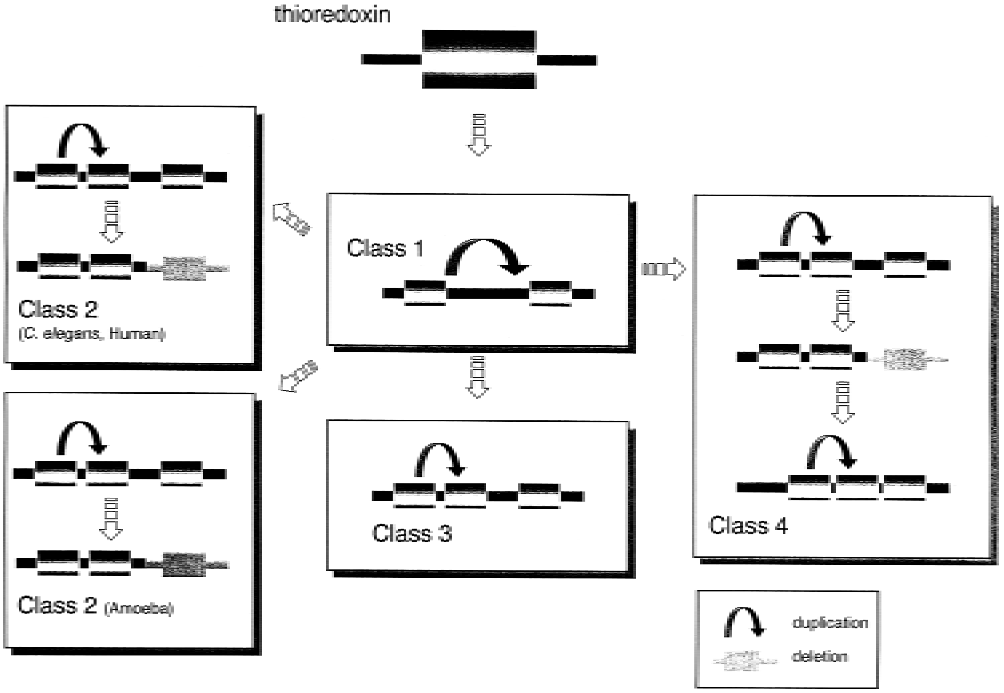

Fig. 3. Inferred evolutionary process of TX domains of PDI. The evolutionary model for class 2 (alfalfa) is not shown (see text).

and their close relationships to the N-terminal domains of class 1. The third subgroup includes the erp5 proteins from humans and *C. elegans.* The evolutionary mechanism of the subgroup was considered to be similar to that of the second subgroup. However, the results shown in category 13 of Table 2, together with the topology shown in Fig. 2, suggested that the two subgroups appeared independently.

Evolutionary Positions of the Class 3 TX Domains

Class 3 proteins have three TX domains in their primary structures. As shown in Fig. 2, the first and second domains formed a cluster in the N-terminal domains of class 1, while the third domains of class 3 formed a cluster in the C-terminal domains of class 1. The tree topology suggests that the class 3 protein evolved from an ancestral class 1 protein by duplication of the N-terminal domain. However, the bootstrap probabilities for the nodes related to the branching of the domains were very low. To verify the significance of the evolutionary scenario, we again applied a series of quartet tests to the data. Table 3 summarizes the results of the quartet tests.

In categories 1, 2, 4, and 5 in Table 3, the first component of each quartet was the first or second domain of a class 3 protein. The second and third components were the N- and C-terminal domains of a class 1 protein. The fourth component was the C-terminal domain of another class 1 protein, which was most distantly related to the third component. All of the results strongly supported the topology ((1, 2), (3, 4)), which suggests that the first and second domains of class 3 belong to the cluster of the class 1 N-terminal domains. Similarly, we examined whether the third domains of class 3 are included in the C-terminal domain cluster of class 1. Categories 3 and 6 in Table 3 show the results of the quartet tests, where the first component of each quartet was the third domain of a class 3 protein. The second and third domains corresponded with the N- and C-terminal domains of a class 1 protein. The fourth component was the N-terminal domain of another class 1 protein, which was most distantly related to the second component. As shown in the table, the topology ((1, 3), (2, 4)) was strongly supported. The topology indicated that the third domains of class 2 belong to the clusters of the C-terminal domains.

The clustering pattern of the first and second domains was also investigated by the quartet tests. Categories 7 and 10 in Table 3 show the results of the tests. The first and second components of each quartet corresponded with the first and second domains of a class 3 protein, and the third and fourth components were the N- and C-terminal domains of a class 1 protein. Only three out of the eight results suggested the ((1, 2), (3, 4)) topology as the ML tree. For the first and third quartets of category 10, the topology was the second best tree with an AIC that was not significantly different from that of the best one. In addition, the tandem duplication of the N-terminal domain seemed to be a simple and probable explanation for the formation of the first and second domains of class 3 proteins. Furthermore, the polypeptide between the second and third domains of the class 3 proteins is similar in size to that between the N- and

Table 3. Results of quartet tests for class 3 PDIs[a]

					Topology					
	Sequence				((1,2),(3,4))		((1,3),(2,4))		((1,4),(2,3))	
Category	1	2	3	4	diff AIC	Boot P	diff AIC	Boot P	diff AIC	Boot P
1	er72human1	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.9013	9.8	0.0920	10.5	0.0067
	↓	er60humanN	er60humanC	bs2trybbC	0.0	0.9426	12.5	0.0095	12.3	0.0479
	↓	pdihumanN	pdihumanC	↓	0.0	0.9621	13.2	0.0105	13.2	0.0274
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.9697	24.5	0.0269	25.0	0.0034
2	er72human2	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.9844	22.8	0.0120	22.9	0.0036
	↓	er60humanN	er60humanC	bs2trybbC	0.0	0.9199	14.7	0.0790	18.0	0.0011
	↓	pdihumanN	pdihumanC	↓	0.0	0.9888	21.1	0.0098	21.1	0.0014
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.9876	34.0	0.0124	36.6	0.0000
3	erp72human3	bs2trybbN	bs2trybbC	pdihumanC	32.0	0.0002	0.0	0.9984	32.0	0.0014
	↓	er60humanN	er60humanC	bs2trybbC	37.0	0.0039	0.0	0.9961	37.3	0.0000
	↓	pdihumanN	pdihumanC	↓	36.6	0.0088	0.0	0.9910	39.1	0.0002
	↓	pdiyeastN	pdiyeastC	↓	20.0	0.0142	0.0	0.9855	20.0	0.0003
4	erp72caeel1	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.9610	14.0	0.0306	14.0	0.0084
	↓	er60humanN	er60humanC	bs2trybbC	0.0	0.8649	11.1	0.1304	13.8	0.0047
	↓	pdihumanN	pdihumanC	↓	0.0	0.9906	24.6	0.0080	24.8	0.0014
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.9847	23.4	0.0135	23.4	0.0018
5	er72caeel2	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.9780	20.5	0.0114	20.5	0.0106
	↓	er60humanN	er60humanC	bs2trybbC	0.0	0.9541	15.9	0.0444	17.1	0.0015
	↓	pdihumanN	pdihumanC	↓	0.0	0.9863	21.8	0.0124	21.8	0.0013
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.9942	34.3	0.0056	35.2	0.0002
6	er72caeel3	bs2trybbN	bs2trybbC	pdihumanC	19.6	0.0014	0.0	0.9839	19.6	0.0147
	↓	er60humanN	er60humanC	bs2trybbN	37.2	0.0024	0.0	0.9967	37.2	0.0009
	↓	pdihumanN	pdihumanC	↓	25.7	0.0323	0.0	0.9662	27.7	0.0015
	↓	pdiyeastN	pdiyeastC	↓	12.2	0.0871	0.0	0.9099	13.3	0.0030
7	er72human1	er72human2	bs2trybbN	bs2trybbC	3.1	0.1457	3.1	0.0817	0.0	0.7726
	↓	↓	er60humanC	er60humanC	5.0	0.1068	0.0	0.7449	4.8	0.1483
	↓	↓	pdihumanN	pdihumanC	0.0	0.4777	2.3	0.0488	0.1	0.4735
	↓	↓	pdiyeastN	pdiyeastC	0.0	0.4950	1.0	0.0888	1.3	0.4162
8	er72human2	er72human3	bs2trybbN	bs2trybbC	42.9	0.0009	0.0	0.9991	42.9	0.0000
	↓	↓	er60humanN	er60humanC	54.6	0.0001	0.0	0.9971	53.3	0.0028
	↓	↓	pdihumanN	pdihumanC	50.8	0.0001	0.0	0.9988	50.8	0.0001
	↓	↓	pdiyeastN	pdiyeastC	39.5	0.0038	0.0	0.9961	39.6	0.0001
9	er72human3	er72human1	bs2trybbN	bs2trybbC	22.9	0.0073	22.9	0.0028	0.0	0.9899
	↓	↓	er60humanN	er60humanC	37.3	0.0016	37.3	0.0004	0.0	0.9980
	↓	↓	pdihumanN	pdihumanC	32.4	0.0010	32.4	0.0034	0.0	0.9956
	↓	↓	pdiyeastN	pdiyeastC	24.5	0.0134	24.5	0.0016	0.0	0.9850
10	er72caeel1	er72caeel2	bs2trybbN	bs2trybbC	0.3	0.4394	1.5	0.0703	0.0	0.4903
	↓	↓	er60humanN	er60humanC	6.6	0.0991	6.7	0.0397	0.0	0.8612
	↓	↓	pdihumanN	pdihumanC	0.9	0.2734	0.0	0.5920	1.0	0.1346
	↓	↓	pdiyeastN	pdiyeastC	0.0	0.7023	3.1	0.0624	2.6	0.2353
11	er72caeel2	er72caeel3	bs2trybbN	bs2trybbC	22.6	0.0082	0.0	0.9903	22.6	0.0015
	↓	↓	er60humanN	er60humanC	65.9	0.0002	0.0	0.9998	65.9	0.0000
	↓	↓	pdihumanN	pdihumanC	32.5	0.0057	0.0	0.9914	32.5	0.0029
	↓	↓	pdiyeastN	pdiyeastC	27.5	0.0166	0.0	0.9815	28.4	0.0019
12	er72caeel3	er72caeel1	bs2trybbN	bs2trybbC	17.4	0.0161	17.4	0.0037	0.0	0.9802
	↓	↓	er60humanN	er60humanC	43.0	0.0004	42.7	0.0022	0.0	0.9974
	↓	↓	pdihumanN	pdihumanC	41.9	0.0000	41.3	0.0041	0.0	0.9959
	↓	↓	pdiyeastN	pdiyeastC	19.9	0.0039	19.8	0.0226	0.0	0.9735

[a] The ''diff AIC'' denotes the difference between the minimal AIC and the AIC of each quartet. Therefore, ''diff AIC'' is 0.0 when the quartet is the topology with the minimal AIC. ''Boot P'' indicates the bootstrap probability of each quartet.

C-terminal domains of the class 1 proteins (see Fig. 1). Figure 3 shows a possible model for the evolution of the class 3 domain structure. Class 3 is considered to have diverged from class 1 by gene duplication. After that, the N-terminal domain was tandemly duplicated, and the current domain structure was acquired. Therefore, the appearance of class 3 is considered to be relatively recent, although it occurred before the species divergence between humans and *C. elegans*. The same evolutionary relationships among class 1 and class 3 were previously suggested by Sahrawy et al. (1996), although the statistical significance was not discussed in their analysis.

Table 4. Results of quartet tests for class 4 PDIs[a]

					Topology					
	Sequence				((1,2),(3,4))		((1,3),(2,4))		((1,4),(2,3))	
Category	1	2	3	4	diff AIC	Boot P	diff AIC	Boot P	diff AIC	Boot P
1	pdirhuman1	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.7985	5.2	0.1814	6.2	0.0201
	↓	er60humanN	er60humanC	bs2trybbC	0.0	0.5543	1.9	0.1082	1.1	0.3375
	↓	pdihumanN	pdihumanC	↓	0.0	0.8989	8.0	0.0338	8.0	0.0673
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.7989	8.0	0.0088	6.5	0.1923
2	pdirhuman2	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.8748	9.1	0.1103	10.0	0.0149
	↓	er60humanN	er60humanC	bs2trybbC	0.0	0.7698	6.4	0.2054	8.4	0.0248
	↓	pdihumanN	pdihumanC	↓	0.0	0.9611	14.8	0.0121	14.8	0.0268
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.9846	21.4	0.0107	21.4	0.0047
3	pdirhuman3	bs2trybbN	bs2trybbC	pdihumanC	0.0	0.9347	11.1	0.0484	11.2	0.0169
	↓	er60humanN	er60humanC	bs2trybbC	0.0	0.7152	5.1	0.2218	6.5	0.0630
	↓	pdihumanN	pdihumanC	↓	0.0	0.8885	6.4	0.0591	6.4	0.0524
	↓	pdiyeastN	pdiyeastC	↓	0.0	0.8569	7.8	0.1387	8.8	0.0044
4	pdirhuman1	pdirhuman2	bs2trybbN	bs2trybbC	0.0	0.5543	1.2	0.0924	0.7	0.3533
	↓	↓	er60humanN	er60humanC	0.0	0.7313	6.4	0.0985	5.8	0.1702
	↓	↓	pdihumanN	pdihumanC	0.0	0.8042	6.4	0.0962	6.4	0.0996
	↓	↓	pdiyeastN	pdiyeastC	0.0	0.8543	6.4	0.0334	6.3	0.1123
5	pdirhuman2	pdirhuman3	bs2trybbN	bs2trybbC	0.0	0.5110	0.3	0.1467	0.3	0.3423
	↓	↓	er60humanN	er60humanC	0.0	0.4249	0.8	0.3098	1.0	0.2653
	↓	↓	pdihumanN	pdihumanC	0.0	0.6685	3.2	0.2977	4.7	0.0338
	↓	↓	pdiyeastN	pdiyeastC	5.7	0.1373	0.0	0.8332	6.0	0.0295
6	pdirhuman3	pdirhuman1	bs2trybbN	bs2trybbC	1.2	0.1008	0.0	0.6997	1.2	0.1995
	↓	↓	er60humanN	er60humanC	0.0	0.6098	3.6	0.2650	4.7	0.1252
	↓	↓	pdihumanN	pdihumanC	0.0	0.6216	5.3	0.0937	3.5	0.2847
	↓	↓	pdiyeastN	pdiyeastC	0.0	0.8139	5.8	0.0258	5.5	0.1603
7	pdirhuman1	pdirhuman2	pdirhuman3	bs2trybbN	0.0	0.7644	4.1	0.0756	4.0	0.1600
	↓	↓	↓	bs2trybbC	0.0	0.5754	2.5	0.0402	1.1	0.3844
	↓	↓	↓	er60humanN	0.0	0.5230	0.8	0.4312	3.2	0.0458
	↓	↓	↓	er60humanC	0.0	0.7574	3.9	0.1696	4.0	0.0730
	↓	↓	↓	pdihumanN	0.0	0.4396	0.3	0.3539	0.5	0.2065
	↓	↓	↓	pdihumanC	0.0	0.7453	5.1	0.1098	4.8	0.1449
	↓	↓	↓	pdiyeastN	0.0	0.5980	2.0	0.3884	4.7	0.0136
	↓	↓	↓	pdiyeastC	0.0	0.8892	8.0	0.0913	8.1	0.0195

[a] The ‘‘diff AIC’’ denotes the difference between the minimal AIC and the AIC of each quartet. Therefore, ‘‘diff AIC’’ is 0.0 when the quartet is the topology with the minimal AIC. ‘‘Boot P’’ indicates the bootstrap probability of each quartet.

Evolutionary Positions of the Class 4 TX Domains

As shown in Fig. 2, the three TX domains of the class 4 protein formed a single cluster. The evolutionary relationships among the three domains are clearly different from those of the class 3 proteins, which also have three TX domains. However, the bootstrap probabilities for the nodes, where the three domains branched, were not very high (49.7% and 36.0%, see Fig. 2). Figure 2 also suggests that all three domains belong to the N-terminal domains of class 1. The three domains are most closely related to the N-terminal domain of trypanosoma BS2, a class 1 protein, although the bootstrap probability for the node connecting these TX domains was also quite low (16.4%). To check the statistical significance of the clustering pattern, we tried the following quartet tests. Table 4 summarizes the results of the tests.

In categories 1 through 3 of Table 4, the first component of each quartet was the first, second, or third domain of a class 4 protein. The second and third components were the N- and C-terminal domains of a class 1 protein, while the fourth component was the C-terminal domain of another class 1 protein, which was most distantly related to the third component. All of the results supported the topology ((1, 2), (3, 4)), which indicates that all three domains belong to the N-terminal domain cluster of class 1.

To examine the relatedness of any pair of the three class 4 domains, a series of quartet tests was designed, as shown in categories 4 through 6 of Table 4. The first and second components of each quartet corresponded to a pairwise combination of the three domains of class 4. The third and fourth components were the N- and C-terminal TX domains of a class 1 protein. As shown in the table, 10 out of the 12 results supported the topology ((1, 2), (3, 4)). The results, together with the tree topology shown in Fig. 2, indicate that any pair of the three domains of class 4 are more closely related to each other than to the N- or C-terminal domains of class 1.

Finally, we checked the statistical significance of the branching order of the three domains of class 4 in category 7 of Table 4. In each quartet test, the first, second, and third components corresponded with the first, second, and third domains of a class 4 protein. The fourth component was the N- or C-terminal domain of a class 1 protein. All of the results supported the topology ((1, 2), (3, 4)). The results, together with tree topology in Fig. 2, suggested that the first domain is more closely related to the second domain than to the third domain.

Considering the results, we can propose a possible explanation for the domain evolution of class 4 (see Fig. 3). Gene duplication of an ancestral class 1 protein yielded another copy of a class 1 PDI gene. Like the cases of class 2 and class 3, the N-terminal domain had been duplicated. Afterward, the C-terminal domain was deleted. The two reinstated domains correspond to the first and the third domains of the current class 4 protein, respectively. Finally, the reinstated N-terminal domain was duplicated again to yield the second domain. This proposed evolutionary scenario explains the clustering of the three domains and their close relatedness to the N-terminal domain of class 1.

Concluding Remarks

We have investigated the evolution of the domain structures of the PDI family. Unexpectedly, our studies suggest that the PDI domain structures have been reorganized independently on many occasions. Through the studies, we found the tendencies of the reorganization. The N-terminal TX domains are apt to be duplicated, while the C-terminal tends to be deleted. At least two domains seem to be required, although the functional meaning is unknown. Further sequence data accumulation would be helpful to examine the significance of the tendencies and the evolutionary meanings.

Here, we have not fully investigated the evolutionary relationships among the four classes, or those within each class. This was mainly because the high sequence divergence and the small number of alignment sites of the TX domains placed these analyses beyond the limits of the current methods for molecular phylogeny. For example, the topology of the N-terminal domains of class 1 differs from that of the C-terminal domains. Further improvement of phylogeny inference from sequence data and the introduction of tertiary structure comparison into molecular phylogeny analysis would clarify the evolutionary relationships among the TX domains.

Note Added at Proof

Recently, two PDIs with a single TX domain were sequenced from yeast. One is MPD1 (Tachikawa et al. 1995, *FEBS Lett* 369:212–216), and the other is MPD2 (Tachikawa et al. 1997, *Biochem Biophys Res Commun* 239:710–714). We tried to identify the evolutionary positions of these TX domains. Then, it was found that TX domain of MPD1 was present in the cluster of the N-terminal TX domains, while that of MPD2 was included in the cluster of C-terminal TX domain.

One of the referees recommended us to state: ‘‘The first step shown in the evolution of Class 2 and 4 (Fig. 3) is a step exactly equivalent to that shown for the origin of Class 3, namely duplication of the N-terminal domain of Class 1. However, it is clearly implied in the figure that Class 3 is not ancestral to Classes 2 and 4.’’

References

Adachi J, Hasegawa M (1995) MOLPHY (programs for molecular phylogenetics) 2.3b.3. Institute of Statistical Mathematics, Tokyo

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Contr 19:716–723

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230

Bourdi M, Demady D, Martin JL, Jabbour Sk, Martin BM, George JW, Pohl LR (1995) cDNA cloning and baculovirus expression of the human liver endoplasmic reticulum P58: characterization as a protein disulfide isomerase isoform, but not as a protease or a carnitine acyltransferase. Arch Biochem Biophys 323:397–403

Cai H, Wang C, Tsou C (1994) Chaperone-like activity of protein disulfide isomerase in the refolding of a protein with no disulfide bonds. J Biol Chem 269:24550–24522

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In; Dayhoff MO (ed) Atlas of protein sequence structure. National Biomedical Research Foundation, Washington, DC, p 345

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Felsenstein J (1993) PHYLIP (phylogeny inference package) version 3.5c. Department of Genetics, University of Washington, Seattle

Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods in enzymology 266:418–427

Freedman RB (1989) Protein disulfide isomerase: multiple roles in the modification of nascent secretory proteins. Cell 57:1069–1072

Freedman RB, Bulleid NJ, Hawkins HC, Paver JL (1989) Role of protein disulphide-isomerase in the expression of native proteins. Biochem Soc Symp 55:167–192

Freedman RB, Hirst TR, Tuite MF (1994) Protein disulphide isomerase: building bridges in protein folding. Trends Biochem Sci 19: 331–336

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12:543–548

Gething M-J, Sambrook J (1992) Protein folding in the cell. Nature 355:33–45

Hayano T, Kikuchi M (1995a) Cloning and sequencing of the cDNA encoding human P5. Gene 164:377–378

Hayano T, Kikuchi M (1995b) Molecular cloning of the cDNA encoding a novel protein disulfide isomerase-related protein (PDIR). FEBS Lett 372:210–214

Higgins DG, Bleasby AJ, Fuchs R (1991) CLUSTAL V: improved software for multiple sequence alignment. Comput Appl Biosci 8:189–191

Hsu MP, Muhich ML, Boothroyd JC (1989) A developmentally regulated gene of trypanosomes encodes a homologue of rat protein-disulfide isomerase and phosphoinositol-phospholipase C. Biochemistry 28:6440–6446

Huang SH, Tomich JM, Wu H, Jong A, Holcenberg J (1991) Human deoxycytidine kinase. Sequence of cDNA clones and analysis of expression in cell lines with and without enzyme activity. J Biol Chem 266:5353

Kishino H, Miyata H, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 31:151–160

Maciukenas M, McCaughey M (1994) TreeTool 2.0.2. Ribosomal RNA Database Project, University of Illinois

Puig A, Gilbert HF (1994) Protein disulfide isomerase exhibits chaperone and anti-chaperone activity in the oxidative refolding of lysozyme. J Biol Chem 269:7764–7771

Rupp K, Birnbach U, Lundstrom J, Van PN, Soling HD (1994) Effects of CaBP2, the rat analog of ERp72, and of CaBP1 on the refolding of denatured reduced proteins. Comparison with protein disulfide isomerase. J Biol Chem 269:2501–2507

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sahrawy M, Hecht V, Lopez-Jaramillo J, Chueca A, Chartier Y, Meyer Y (1996) Intron position as an evolutionary marker of thioredoxins and thioredoxin domains. J Mol Evol 42:422–431

Scherens B, Messenguy F, Gigot D, Dubois E (1992) The complete sequence of a 9,543 bp segment on the left arm of chromosome III reveals five open reading frames including glucokinase and the protein disulfide isomerase. Yeast 8:577–585

Shorrosh BS, Dixon RA (1992) Molecular characterization and expression of an alfalfa protein with sequence similarity to mammalian ERp72, a glucose-regulated endoplasmic reticulum protein containing active site sequences of protein disulphide isomerase. Plant J 2:51–58

Tasanen K, Parkkonen T, Chow LT, Kivirikko KI, Pihlajaniemi T (1988) Characterization of the human gene for a polypeptide that acts both as the beta subunit of prolyl 4-hydroxylase and as protein disulfide isomerase. J Biol Chem 263:16218–16224

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Bonfield J, Burton J, Connell M, Copsey T, Cooper J, et al (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. Nature 368:32–38

Wang C, Tsou C (1993) Protein disulfide isomerase is both an enzyme and a chaperone. FASEB J 7:1515–1517

Wong JM, Bateman E (1994) Cloning of a cDNA encoding an Acanthamoeba castellanii PDI-like protein. Gene 150:175–179