Prediction of the transmembrane regions of β-barrel membrane proteins with a neural network-based predictor

IRENE JACOBONI,[1] PIER LUIGI MARTELLI,[1] PIERO FARISELLI,[1,2] VITO DE PINTO,[3] AND RITA CASADIO[1,2]

[1]Laboratory of Biocomputing, Centro Interdipartimentale per le Ricerche Biotecnologiche (CIRB), Bologna, Italy
[2]Laboratory of Biophysics, Department of Biology, University of Bologna, Bologna, Italy
[3]Laboratory of Biochemistry and Molecular Biology, Department of Chemical Science, University of Catania, Italy

(RECEIVED September 6, 2000; FINAL REVISION January 2, 2001; ACCEPTED January 11, 2001)

Abstract

A method based on neural networks is trained and tested on a nonredundant set of β-barrel membrane proteins known at atomic resolution with a jackknife procedure. The method predicts the topography of transmembrane β strands with residue accuracy as high as 78% when evolutionary information is used as input to the network. Of the transmembrane β-strands included in the training set, 93% are correctly assigned. The predictor includes an algorithm of model optimization, based on dynamic programming, that correctly models eight out of the 11 proteins present in the training/testing set. In addition, protein topology is assigned on the basis of the location of the longest loops in the models. We propose this as a general method to fill the gap of the prediction of β-barrel membrane proteins.

Keywords: Neural networks; secondary structure predictions; multiple sequence alignment; pattern recognition; membrane β strands; prediction of membrane porins

At present, two types of membrane proteins have been characterized: The first includes all the proteins that to a different extent interact with the lipid bilayer of the cytosplamic membrane of all cells (White and Wimley 1999); the second group includes those proteins that during the last ~10 yr have been discovered in the outer membrane of bacteria, chloroplasts, and mitochondria (Schulz 2000). A major distinguishing feature of membrane proteins of the first type is that they span the cytoplasmic membrane with α-helixes, whereas those of the second type interact with the outer membrane with antiparallel β-strands forming barrels, existing as monomers and oligomers (Cowan and Rosenbusch 1994). These chains, referred to as β-barrel membrane proteins (Gouaux 1998; Schulz 2000), comprise the archetypal trimeric porins of Gram-negative bacteria consisting of water-filled channels that nonspecifically mediate the passive transport of ions and small hydrophilic molecules (<6 kD) or select for certain molecules such as malto-oligosaccharides (Schulz 1996). In addition, more recently, other β-barrel membrane proteins have been characterized, and their functions are quite diverse from that of archetypal porins. After the recent atomic resolution of some proteins from enteric bacteria (FepA, Buchanan et al. 1999; and Fuha, Ferguson et al. 1998), it became evident that high-affinity outer membrane receptors that actively translocate large nutrient molecules like iron-siderophore complexes and vitamin B_{12} span the outer membrane with a β-barrel structure. This architecture is also the outer membrane's interacting part of export protein systems for small antibacterial drugs and large protein toxins (TolC in *Escherichia coli*; Koronakis et al. 2000) and that part of the enzyme phospholipase A (OmplA; Snijder et al. 1999) participating in secretion of colicins in *E. coli* and implied in virulence in *Helicobacter pylori*. The structure of Staphyloccal α-hemolysin high-

Reprint requests to: Dr. Rita Casadio, Department of Biology, University of Bologna, Via Irnerio 42, 40126 Bologna Italy; e-mail: casadio@kaiser.alma.unibo.it; fax: 39-051-242576.

Article and publication are at www.proteinscience.org/cgi/doi/10.1110/ps.37201.

lights the fact that the lytic outer transmembrane domain comprises the lower half of a 14-strand antiparallel barrel containing seven homoprotomers (Song et al. 1996). In addition, membrane β-barrels also have been found in OmpA (Pautsch and Schulz 1998) and OmpX (Vogt and Schulz 1999) from *E. coli*, proteins that participate in bacterial conjugation, function as receptors for bacteriophages and colicins, and mediate virulence and pathogenicity. Finally, in eukaryotes, β-barrels are thought to be the functional structure of voltage-dependent anion channels present in the outer membrane of chloroplasts and mitochondria (for review, see Mannella 1998). It seems, therefore, that the β-barrel structure is associated with functions that are more and more relevant to the entire cell metabolism and that are as diverse as active ion transport, passive nutrient intake, membrane anchors, membrane-bound enzymes, and defense against attack proteins. In addition, it is now evident that the different functions are associated with different barrel sizes (ranging from small eight-stranded to large 22-stranded β-barrels) and with different topologies and aggregation number (Schulz 2000).

Although after a decade of analysis the construction principles of β-barrel membrane proteins are known (Sansom and Kerr 1995; Schulz 2000), it is almost impossible to derive three-dimensional models for proteins of the outer membrane. This is because of the fact that unless they belong to the same family, β-barrel membrane proteins share little sequence identity within each other even in the transmembrane spanning regions. It is well documented that in this case, methods based on homology building and threading cannot be successful (Sternberg et al. 1999). It is therefore necessary to be able to locate correctly the transmembrane regions in a sequence to assign the correct barrel topology and eventually build a three-dimensional model on the basis of the existing templates.

This task, however, appears to be more difficult than predicting the topography and topology of all-helical membrane proteins, whose transmembrane domains can be well detected (Jones et al. 1994; Rost et al. 1995, 1996).

When only the archetypal porins were known, it was suggested that strands along the protein sequence could be located using the evaluation of the chemico-physical properties, such as the hydrophobic moment, associated with the transmembrane region (Paul and Rosenbusch 1985; Welte et al. 1991). However, these methods were successful only if used in combination with experimental information (Schirmer and Cowan 1993). Moreover, amphipathicity of β-membrane strands is generally more complex than simple alternating patterns of hydrophobic and hydrophilic residues (Schulz 2000).

Gibbs sampling provides some hints on the alignment of local regions partially overlapping with transmembrane strands (Neuwald et al. 1995), and Hidden Markov models of different porin families can be used to produce alignments that are useful for structure prediction, provided that a given sequence fit the alignment and that a crystallized counterpart is present in the family (Bateman et al. 1999).

An alternative to alignment methods is predictors of β-membrane spanning regions specific for outer-membrane proteins. A rule-based approach for identifying transmembrane β-strands was described and successfully applied to predict a limited number of archetypal porins (Gromiha et al. 1997). More recently, a neural network predictor became available for locating residues along the *Z*-axis of the pores (Diederichs et al. 1998).

In this article, we will use prototypes of the β-barrel membrane proteins crystallized so far for training and testing a neural network-based predictor to locate strands along the protein sequence. The method with a jackknife procedure, using evolutionary information as input, reaches an overall accuracy per residue as high as 78%. In addition, with a model optimization method using a dynamic programming algorithm, eight topological models out of the 11 proteins included in the testing set are correctly predicted. We analyze the results in terms of the network's capability of extracting characteristic features common to the different β-barrel membrane proteins representative of the different barrel architectures (and functions) and propose models for outer membrane proteins not yet solved at atomic level.

Results and Discussion

The database of β*-barrel proteins*

We use a database including 11 β-barrel membrane proteins. They were selected from the PDB database (after clustering the porins and porin-like proteins into homologous families). From each group, we considered one solved structure with a sequence identity <23% to all the other structures in the different families and with the highest crystal resolution within the group. In this way, the database includes 11 proteins. Among these, five porins are active as canonical homotrimers of β-barrels: the integral membrane protein porin from *Rhodobacter capsulatus* (2por; Weiss and Schulz 1992), its counterpart from *Rhodopseudomonas blastica* (1prn; Kreusch and Schulz 1994), the matrix Ompf porin (2omf; Cowan et al. 1995), maltoporin from *Salmonella typhimurium* (2mpr; Meyer et al. 1997), and the sucrose-specific porin ScrY from the same bacterium (1a0s; Forst et al. 1998). The first three porins contain 16 β-strands in the barrel, the second two contain 18 strands. The remaining six proteins of the database act as monomers with one barrel: The outer membrane transporters FepA (1fep; Buchanan et al. 1999) and FhuA (1fcp; Ferguson et al. 1998) from *E. coli* (both with 22 β-strands in the barrel); the integral outer membrane protein X from *E. coli* (OmpX; 1qj8; Vogt and Shulz 1999) and the transmembrane region of the outer membrane protein A from *E. coli* (OmpA;

1bxw; Pautsch and Schultz 1998; both with 8 β-strands in the barrel); the integral membrane phospholipase from *E. coli* (1qd5; Snijder et al. 1999; with 12 β-strands in the barrel); and one subunit of the heptameric transmembrane pore of Staphyloccal α-hemolysin (7ahl; Song et al. 1996; assembling two β-strands to the barrel).

The database, therefore, contains prototypes of all the β-barrel membrane proteins known so far with atomic resolution (Schulz 2000), and in this, it differs from smaller sets of porins used in previous studies (Diederichs et al. 1998). The number of total residues is 3773, 1909 of which are included in 158 β-strands. Figure 1A shows the length distribution of the β-strands of the database and also the different length distribution of the inner and outer loops of the β-barrels in the selected database. It is evident that the length of the transmembrane β-strands ranges from a minimum of six residues to a maximum of 22 residues, with the highest frequency of occurrence centered at 12 residues. Also, it is noticeable that the longest loops in the different barrels are exposed to the external medium (Fig. 1B). These characteristics are taken into account when implementing the predictor (see below).

The predictor at work

A neural network is trained and tested on the selected database with a jackknife procedure. In this way, each protein at the time is tested while the remaining 10 are used to train the network (the level of identity among the different proteins used ranges from 4% to 22% at the most). It is evident (Table 1) that the network performance is significantly improved when the evolutionary profile is used as input, as compared to single sequence. This is particularly true when a nonredundant database of sequences is used to perform the alignment. Sequence profiles are derived using the HSSP files (Dodge et al. 1998) or a program implemented in-

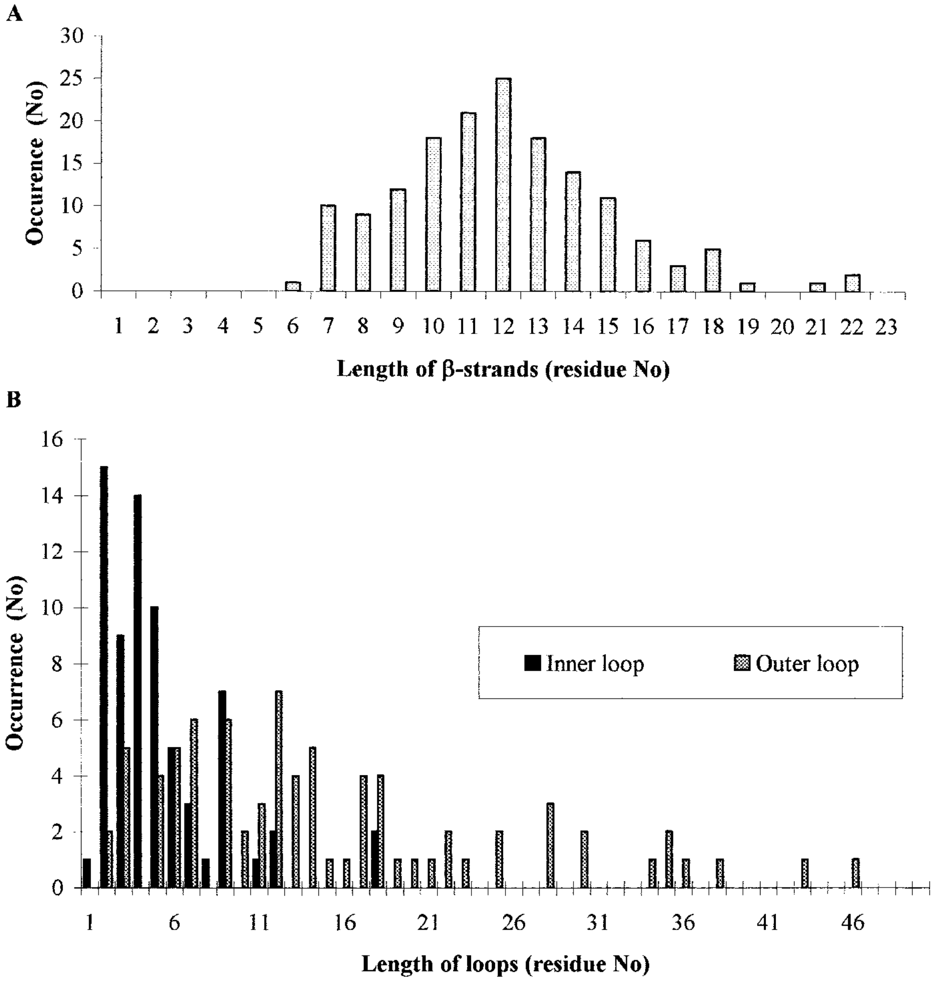

Fig. 1. Characteristic feature of the β-barrels of the outer membrane proteins of the training set. (*A*) Bar plot of the length distribution of the β strands contained in the barrels. (*B*) Bar plot of the length distribution of the inner (black bars) and outer (gray bars) loops of the barrels.

Table 1. *Statistical analysis of the predictive performance*

	Q2	Q(β)	Q(c)	P(β)	P(c)	C(β)	Sov(β)
Training[a]							
Single sequence	0.95	0.94	0.95	0.95	0.94	0.89	0.97
HSSP	0.85	0.86	0.82	0.83	0.85	0.69	0.87
PSI-BLAST	0.89	0.84	0.93	0.92	0.85	0.77	0.91
PHD on β-barrel TM proteins[b]	0.71	0.55	0.88	0.82	0.65	0.45	0.61
PSIpred on β-barrel TM proteins[b]	0.77	0.73	0.83	0.81	0.75	0.56	0.72
Testing[a]							
Single sequence	0.69	0.74	0.64	0.68	0.71	0.38	0.71
HSSP	0.73	0.76	0.70	0.72	0.74	0.46	0.75
PSI-BLAST	0.78	0.74	0.82	0.81	0.76	0.56	0.79

[a] Training and testing of the predictor described in this paper; β = β-strands; c = non β-strands. For the definition of the different statistical indexes, see Materials and Methods.

[b] PHD (cubic.bioc.columbia.edu/predictprotein) and PSIpred (insulin.brunel.ac.uk/psipred) contain respectively 5 and 6 β-barrel TM proteins homologous to those of our selected set in their training sets.

house that computes profiles directly from the alignments performed by PSI-BLAST on a nonredundant database including a number of sequences greater (~500,000) than Swiss-Prot (90,000). For the sake of comparison, the accuracy of the best-performing neural network–based predictors available on the web (PHD and PSIpred) and those using PSI-BLAST is also shown. As described by the authors at their Web sites, PHD and PSIpred were trained on training sets containing a subset of our selected β-barrel membrane proteins. For this reason, we list their performance under "Training" in Table 1. It appears that in spite of this, the three-output predictors are performing with efficiency lower than that of the predictor specific for β barrel membrane proteins, both in training and in testing.

If we consider that our selected database contains ~50% of the residues with β-strand structures, we can estimate that random prediction would give an average accuracy equal to 50%. It is evident from the data shown in Table 1 that even using single sequence as input to the network, the accuracy is 19 points better than random. This value further improves of another 9 points when evolutionary information is extracted from the nonredundant database (for a total of 28 points better than random). The observation that the predictive performance improves when using a large database to extract profiles is in agreement with what was recently observed on the prediction of secondary structure of proteins (Cuff and Barton 2000).

If the base line to score predictive performance is the accuracy of the random predictor, the performance that we obtain in predicting transmembrane β-strands is similar to that of other predictors based on neural networks and developed for the prediction of transmembrane all-helical regions (Rost et al. 1995; Casadio et al. 1996). The rate of false positives (equal to ~1−*P*[β]) is in the range of 20% and is somewhat higher than that noticed before for the all-helical transmembrane domains (16%; Rost et al. 1995), suggesting that transmembrane β-strands are endowed with less representative patterns that the all-helical transmembrane domains.

This finding prompted us to develop an algorithm based on dynamic programming (Needleman and Wunsch 1970) and implementing constraints derived from the models of the transmembrane β-barrel proteins known at atomic resolution (Fig. 1A,B). Network outputs are used to evaluate the score relating the compatibility of a given sequence with a given architectural model of transmembrane β-barrels. For computing this, we rely on a model optimization approach (see Materials and Methods).

This method should, in principle, correct for all the false positives that fall in regions along the protein sequence that do not meet the constraints used to describe β-barrel models present in the database. In Figure 2, two examples of predictions are shown: one is that of the porin chain from *Rhodopesudomonas blastica* (1prn; Fig. 2A), and the other is that of OmpA from *E. coli* (1bxw; Fig. 2B). Network outputs obtained in testing (the protein is not included in the database used for training) are plotted along the protein sequence together with the expected transmembrane β-strands (segments in black), and the results are computed after regularizing outputs with the algorithm based on model optimization (segments in gray). It is evident that both the 16 and 8 β-strands of 1prn and 1bxw are correctly located (see, also, Table 2). This last protein, whose crystal structure became available only recently (Pautsch and Schulz 1998) was predicted with 16 β-strands with other methods mainly relying on comparative modeling (Stathopoulos 1996) and also by a neural network previously described and trained on a set of transmembrane β-barrel proteins smaller than that described in this article (Diederichs et al. 1998). This last predictor was trained using single sequence and provides only network outputs, without minimizing false positives (http://strucbio.biologie.uni-konstanz.de/~kay).

All the models predicted with our procedure (outlined above) are presented in Table 2 and compared to the expected structures. Of the β-strands, 93% are correctly located, and eight out of 11 models are also correctly assigned (73%). Evidently, in some cases the model optimization algorithm is not sufficient to cancel out false positives (the presence of a β-strand in a wrong position in the sequence), or alternatively, network outputs are not enough strong to originate a transmembrane segment (the absence of a β-strand in the correct position). Our predictor fails in correctly locating one transmembrane strand in 1fcpA, 2mprA and 2por. In all the remaining proteins of the testing set, transmembrane β-strands are correctly located, although

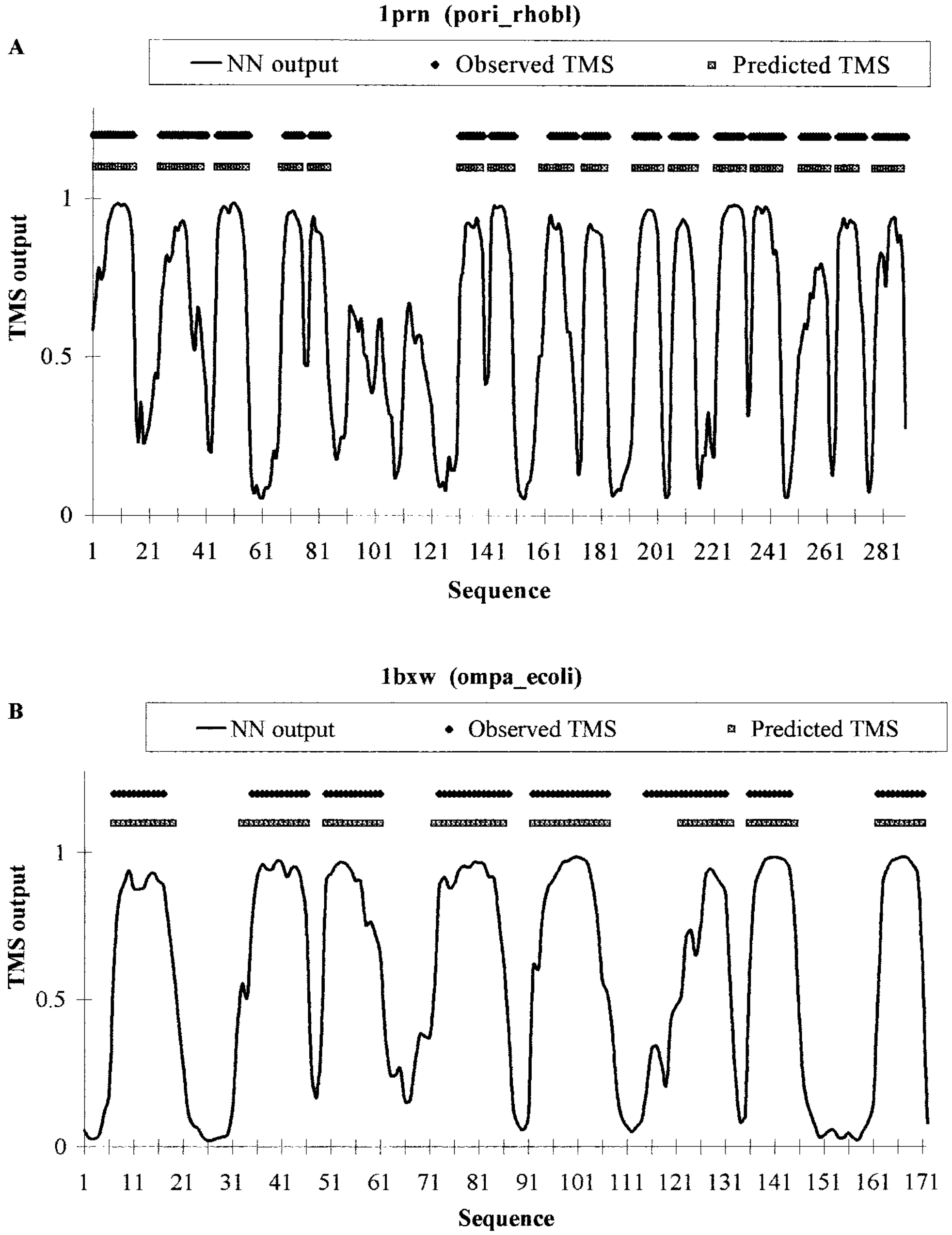

Fig. 2. The predictor at work. Outputs of the neural network predictor are plotted along the protein sequence and originate a pattern whose peaks correspond to regions of high propensity for the membrane β strand structure. A model optimization algorithm based on dynamic programming (see Materials and Methods) selects the optimal model for a given sequence (gray segments) using constraints derived from the actual model of the β barrel membrane proteins in the training set. The optimal model is compared to the observed model (black segments). (*A*) Porin from *Rhodobacter capsulatus* (1prn). (*B*) Outer membrane protein A (OMPA) from *Escherichia coli* (1bxw).

with not a perfect overlapping with the observed corresponding regions.

After selecting the optimal model, the predictor gives also the protein topology. In the case of membrane proteins, topology refers to the protein organization with respect to the membrane phase. On the basis of the observation that in bacterial porins the longest loops are facing the extracellular space, topology is assigned after computing which side of the barrel is endowed with the longest loops. In this way, and considering the predicted models listed in Table 2, the topology of all the 11 proteins of the database is also correctly assigned. It should be noticed, however, that this rule might not hold for β-barrel membrane proteins of mitochondria and chloroplasts, for which data are not yet available.

Table 2. *Observed and predicted transmembrane β-strands for the selected data base of β-barrel transmembrane proteins*

Protein	Observed TMS*	Predicted TMS*	Protein	Observed TMS*	Predicted TMS*	Protein	Observed TMS*	Predicted TMS*
			1fep	235-249	235-247	2mprA	2-14	2-13
1a0sP	3-16	3-14	*(continued)*	272-286	272-286	*lamb_ecoli*	40-47	40-51
scry_salty	47-58	47-58		292-306	292-305		57-68	57-69
	64-75	63-75		317-335	322-335		80-89	81-89
	89-98	89-98		339-354	339-354		99-105	99-108
	111-118	111-120		367-382	368-382		124-130	-
	136-143	132-137		387-398	388-398		138-153	136-147
	152-163	152-161		402-414	408-414		-	152-158
	171-183	172-180		419-430	418-429		164-177	172-179
	186-194	186-195		461-474	461-474		185-197	186-196
	216-226	216-232		477-489	477-489		211-223	211-223
	236-246	236-249		508-515	508-513		226-236	226-236
	265-274	265-273		534-548	533-548		271-284	271-281
	281-293	281-293		561-569	562-570		288-301	288-299
	301-314	301-315		575-584	575-587		307-319	307-319
	319-333	318-331		610-620	611-620		324-337	324-337
	343-356	343-356		625-632	623-632		343-356	344-356
	369-379	369-380		671-679	671-679		366-378	366-380
	402-412	401-412					407-420	410-420
			1prn	1-15	2-15			
1bxwA	7-17	7-19	*pori_rhobl*	25-41	25-39	2omf	9-23	10-24
ompa_ecoli	35-46	33-46		45-56	45-55	*ompf_ecoli*	40-50	39-50
	50-61	50-61		69-75	68-75		55-66	55-66
	73-87	72-86		78-84	78-84		80-90	79-91
	92-107	92-107		131-139	131-139		94-100	95-108
	115-131	122-132		142-150	142-149		136-141	134-143
	136-144	136-145		163-172	160-170		151-158	148-158
	162-171	162-171		175-183	175-182		173-182	174-182
				193-201	193-202		185-195	185-195
1fcpA	143-151	144-150		206-214	206-214		210-222	211-220
fhua_ecoli	155-165	155-160		222-232	222-232		225-235	225-237
	172-184	170-184		234-245	235-244		253-264	261-265
	191-202	-		252-261	252-261		269-281	276-280
	209-220	209-218		265-274	265-271		289-302	291-295
	256-271	256-271		278-288	278-288		307-316	311-315
	276-296	276-292					331-339	335-339
	325-348	321-329	1qd5A	27-33	26-34			
	-	332-347	*pal_ecoli*	53-64	53-65	2por	1-15	2-15
	353-374	353-367		74-86	74-89	*pori_rhoca*	18-35	19-34
	411-432	418-432		99-111	99-112		39-46	39-49
	436-453	436-448		123-132	115-131		59-65	59-70
	459-475	467-477		142-154	142-147		68-74	-
	481-492	481-491		157-166	150-166		-	86-93
	507-518	511-518		185-191	178-187		118-125	118-124
	525-542	525-538		194-202	188-204		128-135	127-134
	554-567	557-568		209-218	209-217		148-158	148-158
	573-588	573-586		224-233	222-237		161-171	161-170
	603-613	603-613		243-252	240-253		181-192	179-191
	621-630	621-632					195-206	195-201
	646-656	641-654	1qj8A	2-14	3-13		227-240	228-240
	666-672	657-673	*ompx_ecoli*	21-31	24-32		243-254	243-253
	696-704	695-704		37-51	39-51		258-271	260-270
				57-71	58-66		275-285	275-279
1fep	144-154	144-151		77-94	77-93		292-300	292-300
fepa_ecoli	161-170	162-169		98-115	104-115			
	177-188	177-187		121-132	121-131	7ahlA	7-18	6-17
	217-229	225-230		135-147	135-147	*hla_staav*	32-43	35-44

* TMS = Transmembrane β-strands.

Predicting other β-barrel transmembrane proteins

While this work was in progress, the crystal of TolC became available (Koronakis et al. 2000). The complex has a channel–tunnel structure spanning the region from the outer membrane up to the inner membrane and is assembled as a trimer of 428-residue identical protomers. Spanning the outer membrane, the protomers form a transmembrane β-barrel of 12 β-strands. We tested one protomer in the transmembrane region, and the results are listed in Table 3. It is evident that our predictor correctly locates the four transmembrane β-strands of the protomer.

The second prediction that we show is that of Omp32, an anion selective porin from *Comamonas acidovorans*, whose crystal structure is announced for June 2001 (Zeth et al. 2000) and was already a target (target 70) during the last CASP3 competition (Orengo et al. 1999). The sequence shows a 23% level of identity with ompf_ecoli, and as soon

Table 3. *Prediction of β-barrel transmembrane β-regions of TolC and OMP32*

Protein	Observed TMS	Predicted TMS
lek9	41–52	41–51
tolc__ecoli	61–74	64–74
	247–262	247–255
	280–289	282–289
OMP32[a]	2–16	2–9
om32_comac	—	12–29
	36–46	—
	52–62	49–63
	73–83	72–78
	86–92	81–93
	137–145	133–140
	148–155	148–155
	171–180	167–177
	183–193	182–195
	202–212	199–213
	218–227	216–232
	237–248	237–250
	254–264	253–262
	269–281	275–292
	286–295	—
	—	295–312
	323–331	322–331

[a] Observed structure of OMP32 refers to the model T0070TS108_1 submitted to the CASP3 competition by Fidelis' group (Venclovas et al. 1999).

as it will be available, it can be included in the database. At present, we tested our results by comparing them with the model that scored the less root mean square deviation (RSMD = 0.35 nm; http://predictioncenter.llnl.gov/casp3/results) to the crystal structure (Venclovas et al. 1999). As shown in Table 3, our predictor indicates a putative 16-stranded transmembrane β-barrel, with the location of 14 out of 16 strands in agreement with the computed model. These results confirm that the predictor is endowed with a generalization capability sufficient to predict with good accuracy transmembrane segments even in proteins distantly related to those of the training set.

Is the predictor of transmembrane β-barrel proteins necessary?

At present, the accuracy of secondary structure prediction is quite high (Cuff and Barton 1999). So one may wonder to which extent neural networks are capable of capturing and generalizing the characteristic features of transmembrane β-strands as compared to those of globular ones.

We trained and tested by cross validation a two-output network for discriminating globular β-strands from coil structure in all-β globular proteins. The overall accuracy of the network reaches 74% with a correlation coefficient equal to 0.49. When, with this network, the transmembrane β-strands are predicted, the accuracy is 69%, compared with that of 78% obtained when the specific network is used on the same testing set (Table 4). This tells us that the predictor trained on β-barrel transmembrane proteins captures distinguished features of transmembrane β-strands that are not included in the same structural type of globular proteins.

Table 4. *Efficiency of different predictors on transmembrane and globular β-strands*

	Q2	Q(β)	Q(c)	P(β)	P(c)	C(β)	Sov(β)
Trained on β-barrel TM proteins Tested on β-barrel TM proteins	0.78	0.74	0.82	0.81	0.76	0.56	0.79
Trained on all-β proteins Tested on all-β proteins	0.74	0.71	0.78	0.78	0.71	0.49	0.78
Trained on β-barrel TM proteins Tested on all-β proteins	0.63	0.50	0.76	0.68	0.59	0.27	0.68
Trained on all-β proteins Tested on β-barrel TM proteins	0.69	0.59	0.79	0.74	0.65	0.38	0.65

The all-β set (β-strand > 45%, α-helix < 5%; Zhang and Chou 1992) contains 59 proteins extracted from the PDB_select, release June 1998.

In addition, we may ask whether different functions are requiring distinct features. We divided our testing (and training set) in two subsets containing, respectively, trimeric porins and monomeric β-barrel ones, performing different functions (see above). Each subset was used to train a network, and the other was predicted. The results (Table 5)

Table 5. *Predicting β-barrel transmembrane proteins with different functions*

	Q2	Q(β)	Q(c)	P(β)	P(c)	C(β)	Sov(β)
Trained on porins Tested on porins	0.79	0.79	0.79	0.82	0.76	0.58	0.79
Trained on other β-barrel TM proteins Tested on other β-barrel TM proteins	0.76	0.70	0.81	0.77	0.75	0.52	0.72
Trained on porins Tested on other β-barrel TM proteins	0.77	0.80	0.75	0.75	0.80	0.55	0.82
Trained on other β-barrel TM proteins Tested on porins	0.80	0.78	0.82	0.83	0.76	0.60	0.78

Porins: 1a0sP, 1prn, 2omf, 2mprA, 2por.
Other β-barrel TM proteins: 1bxwA, 1fcpA, 1fep, 1qd5A, 1qj8A, 7ahlA.

indicate that the predictive efficiency is rather similar in both cases, suggesting that β-barrel architectures are endowed with the same characteristic patterns independent of the function. This adds to the network capability of extracting general features common to all the β-barrel transmembrane proteins.

Conclusions

We propose the use of the predictor described in this work to locate putative transmembrane β-strands in β-barrel-containing membrane proteins. This method may help to build the three-dimensional model of β-barrel membrane proteins by threading on templates of similar architecture.

Our predictor implements an algorithm based on model optimization, which selects network predictions on the basis of the transmembrane β-barrel architectures presently known at atomic resolution. For this and for using evolutionary information, it is presently the only one that is implemented and based on neural network that is capable of correctly assigning 93% of the transmembrane β-strands known at atomic resolution. In addition, our analysis highlights that a neural network is capable of capturing features that are characteristic of transmembrane β-strands, as compared to globular ones, and that these features are shared by the transmembrane β-barrels performing different functions. It is therefore feasible that when new examples will be known at atomic resolution, this method will be potentiated.

The predictor is presently available on request at http://www.biocomp.unibo.it.

Materials and methods

The neural network–based predictor

A feed-forward neural network is implemented and trained with the back-propagation algorithm (Rumelhart et al. 1986) to discriminate membrane β-strands from extra membrane regions in the β-barrel membrane proteins of the database. The network architecture basically consists of perceptrons with one hidden layer containing five hidden nodes and an input window spanning nine residues. Two output nodes are considered ("β" and "not β"). The architecture of the predictor is extended to include a second cascaded network to filter out spurious assignments. Other network architectures (a smaller or greater number of neurons in the hidden layer) and lengths of the input window (from five to 15) were also tried, and the one described above was found to give the best predictive performance.

Evolutionary information is given as input in the form of sequence profiles after multiple sequence alignments. Sequence alignments were derived from the HSSP database (Dodge et al. 1998) in which alignments were constructed using BLAST (Altschul et al. 1990) to search the sequence database and MAXHOM (Sander and Schneider 1991) to align the sequences. Moreover, we used PSI-BLAST (Altschul et al. 1997; one round with threshold equal to 0.001) to search a nonredundant database (available at http://www.ncbi.nlm.nih.gov/BLAST). We generated sequence profiles from its outputs by means of a newly implemented program. This is based on the notion that the PSI-BLAST complete outputs contain the local pairwise alignments of the query sequence with all the extracted sequences. From this, it is possible to compute a profile by merging each local pairwise alignment.

β-barrel membrane proteins taken from the PDB database were clustered into different homology groups using CLUSTALW (Thompson et al. 1994).

Selecting the model

An algorithm based on dynamic programming uses the network outputs to locate the transmembrane β-strands along the protein sequence by model optimization. A similar algorithm was previously used to locate transmembrane α-helices (Jones et al. 1994). The one we implement takes advantage of the notion that transmembrane β-strands in the prototypes of β-barrel membrane proteins are even in number and range from two to 22 in the sequence (Fig. 1). A recursive algorithm generates a scoring matrix for each predicted sequence by evaluating the total sum of the output differences along a segment of fixed length. Minimal and maximal lengths are derived from the database of selected proteins (Fig. 1A). A model is selected by evaluating the optimal score among those satisfying the observed constraints in the crystals.

For a given sequence position j and for a given model i (i is the number of β-strands), the scoring matrix **S** is computed as

$$S^{i}{}_{j} = \max_{l=\beta_{min} \rightarrow \beta_{max}} \{s^{l}{}_{j} + \max_{k=j+l+L \rightarrow n} \{S^{i-1}{}_{k}\}\} \tag{1}$$

where L and n are the minimum length of a loop segment and the protein length, respectively; $s^{l}{}_{j}$ is the score associated with a transmembrane strand of length l at position j in the sequence.

Topology is then predicted by simply comparing the length of the loops of the two sides of the barrel and labeling as extracellular the barrel side with the longest loops.

Scoring the prediction

The efficiency of the predictors is scored using the statistical indexes defined in the following.

The network accuracy is

$$Q2 = P/N \tag{2}$$

where P is the total number of correct membrane β-strand predictions and N is the total number of possible predictions.

The correlation coefficient C is defined as

$$C(\beta) = (p(\beta)*n(\beta) - u(\beta)*o(\beta))/[(p(\beta) + u(\beta))(p(\beta) + o(\beta))(n(\beta) + u(\beta))(n(\beta) + o(\beta))]^{1/2} \tag{3}$$

where, for each class β, p(β), and n(β) are, respectively, the total number of correct predictions and correctly rejected assignments, whereas u(β) and o(β) are the numbers of under and over predictions.

The accuracy for each discriminated structure s is evaluated as

$$Q(\beta) = p(\beta)/[p(\beta) + u(\beta)] \tag{4}$$

where p(β) and u(β) are the same as in Equation (3).

The probability of correct predictions $P(\beta)$ is computed as

$$P(\beta) = p(\beta)/[p(\beta) + o(\beta)] \quad (5)$$

where $p(\beta)$ and $o(\beta)$ are the same as in Equation 3.

The segment-based measure (Sov) of the assessment of transmembrane β-strands is computed as previously described (Zemla et al. 1999).

Acknowledgments

This work was partially supported by a grant of the Ministero della Universitá e della Ricerca Scientifica e Tecnologica (MURST) for the project "Structural, Functional and Applicative Prospects of Proteins from Termophiles," and by a grant for a target project in Biotechnology of the Italian Centro Nazionale delle Ricerche (CNR), both delivered to R.C.

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

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acid Res.* **25:** 3389–3402.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., and Sonnhammer, E.L. 1999. Pfam 3.1: 1313 multiple alignments match the majority of proteins. *Nucleic Acids Res.* **27:** 260–262.

Buchanan, S.K., Smith, B.S., Venkatramani, L., Xia, D., Esser, L., Palnitkar, M., Chakraborty, R., van der Helm, D., and Deisenhofer, J. 1999: Crystal structure of the outer membrane active transporter FepA from *Escherichia coli*. *Nat. Struct. Biol.* **6:** 56–63.

Casadio, R., Taroni, C., Fariselli, P., and Compiani, M. 1996 A predictor of transmembrane alpha-helix domains of proteins based on neural networks. *Eur. Biophys*.

Cowan, S.W. and Rosenbusch, J.P. 1994. Folding pattern diversity of integral membrane proteins. *Science* **264:** 914–916.

Cowan, S.W., Garavito, R.M., Jansonius, J.N., Jenkins, J.A., Karlsson, R., Konig, N, Pai E.F., Pauptit, R.A., Rizkallah, P.J., Rosenbusch, J.P., et al. 1995. The structure of OmpF porin in a tetragonal crystal form. *Structure* **3:** 1041–1050.

Cuff, J.A. and Barton, G.J. 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40:** 502–511.

Diederichs, K., Freigang, J., Umhau, S., Zeth, K., and Breed, J. 1998. Prediction by neural network of outer membrane β-strand protein topology. *Protein Sci.* **7:** 2413–2420.

Dodge, C., Schneider, R., and Sander, C.1998. The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acid Res.* **26:** 313–315.

Ferguson, A.D., Hofmann, E., Coulton, J.W., Diederichs, K., and Welte, W. 1998. Siderophore-mediated iron transport: Crystal structure of FhuA with bound lipopolysaccharide. *Science* **282:** 2215–2220.

Forst, D., Welte, W., Wacker, T., and Diederichs, K. 1998. Structure of the sucrose-specific porin ScrY from *Salmonella typhimurium* and its complex with sucrose. *Nat. Struct. Biol.* **5:** 37–46.

Gouaux, E. 1998. Roll out the barrel *Nat. Struct. Biol.* **5:** 931–932.

Gromiha, M.M., Majumdar, R., and Ponnuswamy, P.K. 1997. Identification of membrane spanning β strands in bacterial porins. *Protein Engin.* **10:** 497–500.

Jones, D.T., Taylor, W.R., Thornton, J.M. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33:** 3038–3049.

Koronakis, R., Sharff, A., Koronakis, E., Luisi, B., and Hughes, C. 2000. Crystal structure of the bacterial membrane protein To1C central to multidrug efflux and protein export. *Nature* **405:** 914–919.

Kreusch, A. and Schulz, G.E. 1994: Refined structure of the porin from *Rhodopseudomonas blastica*: Comparison with the porin from *Rhodobacter capsulatus*. *J. Mol. Biol.* **243:** 891–905.

Mannella, C.A. 1998. Conformational changes in the mitochondrial channel protein, VDAC, and their functional implications. *J. Struct. Biol.* **121:** 207–218.

Meyer, J.E., Hofnung, M., and Schulz, G.E. 1997. Structure of maltoporin from *Salmonella typhimurium* ligated with a nitrophenyl-maltotrioside. *J. Mol. Biol.* **266:** 761–775.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.

Neuwald, A.E., Liu, J.S., and Lawrence, C.E. 1995. Gibbs sampling: Detection of bacterial outer membrane protein repeats. *Protein Sci.* **4:** 1618–1632.

Orengo, C.A., Bray, J.E., Hubbard, T., LoConte, L., and Sillitoe, L. 1999. Analysis and assessment of ab initio three-dimensional prediction. secondary structure and contacts predictions. *Proteins Suppl.* **3:** 149–170.

Paul, C. and Rosenbusch, J.P. 1985. Folding patterns of porin and bacteriorhodopsin. *EMBO J.* **4:** 1594–1597.

Pautsch, A. and Schulz, G.E. 1998. Structure of the outer membrane protein A transmembrane domain. *Nat. Struct. Biol.* **5:** 1013–1017.

Rost, B., Casadio, R., Fariselli, P., and Sander, C. 1995. Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4:** 521–533.

Rost, B., Fariselli, P., and Casadio, R. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5:** 1704–1718.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. 1986. Learning representation by back-propagating errors. *Nature* **323:** 533–536.

Sander, C. and Schneider, R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9:** 56–68.

Sansom, M.S.P. and Kerr, I.D. 1995. Transbilayer pores formed by β-barrels: molecular modeling of pore structures and properties. *Biophys. J.* **69:** 1334–1343.

Schirmer, T. and Cowan, S.W. 1993. Prediction of membrane spanning B-strands and its application to maltoporin. *Protein Sci.* **2:** 1361–1363.

Schulz, G.E. 1996. Porins: General to specific, native to engineered passive pores. *Curr. Opin. Struct. Biol.* **6:** 485–490.

Schulz, G.E. 2000. β-barrel membrane proteins. *Curr. Opin. Struct. Biol.* **10:** 443–447.

Snijder, H.J., Ubarretxena-Belandia, I., Blaauw, M.I., Kalk, K.H., Verheij, H.M., Egmond, M.R., Dekker, N., and Dijkstra, B.W. 1999. Structural evidence for dimerization-regulated activation of an integral membrane phospholipase. *Nature* **401:** 717–721.

Song, L., Hobaugh, M.R., Shustak, C., Cheley, S., Bayley, H., and Gouaux, J.E. 1996. Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore. *Science* **274:** 1859–1866.

Stathopoulos, C. 1996. An alternative topological model for *Escherichia coli* OmpA. *Protein Sci.* **5:** 170–173.

Sternberg, M.J., Bates, P.A., Kelley, L.A., and MacCallum, R.M. 1999. Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.* **9:** 368–373.

Thompson, J.D., Higgins, D.J., and Gibson, T.J. 1994. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acid Res.* **22:** 4673–4680.

Venclovas, C., Ginalski, K., and Fidelis, K. 1999. Addressing the issue of sequence-to-structure alignments in comparative modeling of CASP3 target proteins. *Proteins Suppl.* **3:** 73–80.

Vogt, J. and Schulz, G.E. 1999. The structure of the outer membrane protein Ompx from *Escherichia coli* reveals mechanisms of virulence. *Structure* **7:** 1301–1309.

Welte, W., Weiss, M.S., Nestel, U., Weckesser, J., Schiltz, E., and Schulz, G.E. 1991. Prediction of the general structure of OmpF and PhoE from the sequence and structure of porin from *Rhodobacter capsulatus*: Orientation of porin in the membrane. *Biochim. Biophys. Acta* **1080:** 271–274.

Weiss, M.S. and Schulz, G.E. 1992. Structure of porin refined at 1.8 Å resolution. *J. Mol. Biol.* **227:** 493–509.

White, S.H. and Wimley, W.C. 1999. Membrane protein folding and stability: Physical principles. *Annu. Rev. Biophys. Biomol. Struct.* **28:** 319–365.

Zemla, A., Venclovas, C., Fidelis, K., and Rost, B. 1999. A modified definition of Sov, a segment-based measure of protein secondary structure prediction assessment. *Proteins* **34:** 220–223.

Zeth, K., Diederichs, K., Welte, W., and Engelhardt, H. 2000. Crystal structure of Omp32, the anion-selective porin from *Comamonas acidovorans*, in complex with the a periplasmic peptide at 2.1 Å resolution. *Structure* **8:** 981–992 (PDB code 1E54).

Zhang, C.T. and Chou, K.C. 1992. An optimization approach to protein structural class from amino acid composition. *Protein Sci.* **1:** 401–408.