The EMBO Journal vol.5 no.11 pp.3021-3027, 1986

The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology

Gunnar von Heijne

Research Group for Theoretical Biophysics, Department of Theoretical Physics, Royal Institute of Technology, S-100 44 Stockholm, Sweden

Communicated by K.Simons

The amino acid distribution in membrane spanning segments and connecting loops in bacterial inner membrane proteins was analysed. The basic residues Arg and Lys are four times less prevalent in periplasmic as compared to cytosolic connecting loops, whereas no comparable effect is observed for the acidic residues Asp and Glu. Also, Pro is shown to be tolerated to a much larger extent in membrane spanning segments with their N-terminus pointing towards the cytosol than in those with the opposite orientation. The significance of these findings with regard to the mechanism of biogenesis of bacterial inner membrane proteins is discussed.

Key words: membrane protein topology/hydropathy plot/structure prediction/inner membrane protein

Introduction

Many membrane proteins such as receptors, pore-forming proteins, ion pumps, nutrient and metabolite transporters, and photosynthetic proteins are absolutely essential for the cell's communication with the outside world. Over the past few years the primary sequences of a large number of membrane proteins have been determined, and the importance of long hydrophobic segments that presumably span the membrane as helices has been recognized (reviewed in von Heijne, 1985). However, the sequence characteristics that determine the trans-membrane topology of the protein still remain largely unknown.

The ability to predict (i) the topology, and (ii) the fully folded structure from the primary sequence is currently limited by our ignorance of the way membrane proteins are inserted into their target membrane. 'Simple' membrane proteins with only one hydrophobic membrane spanning segment generally are made with an N-terminal signal sequence that somehow initiates translocation of the first part of the chain through the membrane. The hydrophobic trans-membrane segment presumably halts this translocation process and anchors the protein to the membrane in its final topology.

The biogenesis of membrane proteins of the 'complex' variety (i.e. proteins with multiple hydrophobic spanning segments) on the other hand can be envisaged as proceeding via two mechanistically very different routes: either by a sequential 'threading' back and forth across the membrane starting from the most N-terminal spanning segment — the topology thus being determined by a succession of 'start' and 'stop' signals (Blobel, 1980) — or by an insertion mechanism where neighbouring hydrophobic segments pair up and penetrate the membrane as 'helical hairpins' (Engelman and Steitz, 1981). In the first model, the topology is essentially determined by the hydrophobic spanning segments alone, whereas the topology in the helical hairpin model is an outcome of kinetic competition between the formation and insertion of all the possible membrane spanning nearest-neighbor pairs in a process where the characteristics of the connecting segments between the hydrophobic stretches should be decisive.

Interestingly, in the recently determined X-ray structure of the photosynthetic reaction centre from *Rhodopseudomonas viridis* (Deisenhofer *et al.*, 1985) a striking charge asymmetry across the membrane has been observed (Michel *et al.*, 1986) with the periplasmic connecting segments or loops generally being more negative than the cytosolic loops. It will now be shown that such a charge asymmetry seems to be common to all bacterial inner membrane proteins, that it results from a bias in the distribution of positively (Arg and Lys) but not negatively (Asp and Glu) charged residues, and that the distribution of positively charged residues in connecting loops may be used to aid in the prediction of the trans-membrane topology of 'complex' bacterial inner membrane proteins.

Results and Discussion

Arg and Lys (but not Asp and Glu) are four times less prevalent in periplasmic relative to cytosolic loops

As described under Materials and methods, a number of bacterial inner membrane proteins in addition to the *Rps. viridis* reaction centre complex have been sufficiently well characterized in terms of their trans-membrane topology to serve as a reasonably reliable database. These proteins are listed in Table I, together with their assumed topologies and a specification of the membrane spanning and connecting segments included in the statistical analysis.

Amino acid counts were collected for four samples, namely periplasmic and cytosolic loops of length 65 residues or less, and membrane spanning segments with their N-terminus facing the cytosolic and periplasmic side of the membrane, Table II. A highly significant difference ($P < 0.001$) in the incidence of postively charged residues was found between the periplasmic and cytosolic loops: $f_{Arg+Lys} = 4.2\%$ in the periplasmic loops versus 15.8% in the cytosolic connecting segments, almost a 4-fold difference. In a control sample of 72 soluble cytosolic bacterial proteins, $f_{Arg+Lys} = 12.3\%$, and in a sample of 45 soluble periplasmic or extracellular bacterial proteins $f_{Arg+Lys} = 10.0\%$ (see Methods). Pro seems to be enriched in the periplasmic loops ($P < 0.001$), with $f_{Pro} = 8.5\%$ versus 3.8%, 4.1% and 4.2% for the cytosolic loops, soluble cytosolic and soluble periplasmic proteins respectively. There is also a marginally significant 1.6-fold reduction ($P < 0.025$) in negatively charged residues in the periplasmic loops where $f_{Asp+Glu} = 7.2\%$ versus 11.2%, 12.8% and 11.8% for the cytosolic loops, soluble cytosolic, and soluble periplasmic proteins.

Not surprisingly, both spanning segment samples are enriched about 2-fold for hydrophobic residues ($f_{Phe+Ile+Leu+Met+Val} = 55\%$) compared with the loops and soluble proteins; more interestingly, Pro is significantly reduced ($P < 0.01$) only in those spanning segments that have their N-terminus towards the periplasmic side of the membrane ($f_{Pro} = 0.9\%$ versus ~4% for the other samples).

In an attempt to extend these calculations to a larger set of proteins, a total of 66 bacterial inner membrane proteins were collected from the National Biomedical Research Foundation (NBRF) Protein Sequence Database (Release 7.0) and from the literature, see Materials and methods. In a preliminary step, hydrophobicity analysis was carried out as described under Materials and methods (essentially, each sequence was partitioned into non-overlapping 23-residue segments starting from the most hydrophobic 19-residue segment and working downwards), and the distribution of peak hydrophobicity values was determined for this sample as well as for the sample of soluble cytosolic proteins. The results are presented in Figure 1, where the bimodal

Table I. Bacterial inner membrane proteins. The topology is indicated by showing the number of positive and negative residues in each connecting loop in its proper cytosolic or periplasmic location, starting from the N-terminus. Loops in square brackets are not included in the amino acid statistics since they are longer than 65 residues

Protein	Spanning segments	Topology		Reference
		cyt	per	
A. Proteins with well-characterized topology				
Reaction centre				
H-subunit	12–30		N–(+0/–2)	Michel *et al.* (1985)
		[+30/–36]–C		
L-subunit	30–48	N–(+3/–2)		Michel *et al.* (1986)
	84–102		(+1/–2)	
	113–131	(+3/–2)		
	177–195		(+1/–1)	
	232–250	(+5/–4)		
			(+0/–4)–C	
M-subunit	53–71	N–(+2/–6)		Michel *et al.* (1986)
	111–129		(+2/–4)	
	148–166	(+3/–1)		
	206–224		(+1/–2)	
	266–284	(+6/–6)		
			(+0/–2)–C	
Bacteriorhodopsin	24–42		N–(+1/–3)	Dunn *et al.* (1981)
	57–75	(+3/–2)		
	96–114		(+1/–2)	
	121–139	(+0/–4)		
	151–169		(+2/–0)	
	190–212	(+4/–2)		
	219–237		(+0/–3)	
		(+3/–6)–C		
Light-harvesting complex				
LH1	23–41	N–(+4/–1)		Drews (1985)
			(+1/–1)–C	
LH2	27–45	N–(+1/–5)		Drews (1985)
			(+1/–1)–C	
malF	17–35	N–(+4/–2)		Froshauer and Beckwith (1984)
	40–58		(+0/–1)	
	73–91	(+3/–0)		
	277–295		[+18/–21]	
	319–337	(+3/–1)		
	371–389		(+3/–2)	
	418–436	(+3/–4)		
	486–504		(+2/–4)	
		(+3/–2)–C		
M13 coat protein precursor	5–23	N–(+3/–0)		van Wezenbeek *et al.* (1980)
	46–64		(+1/–4)	
		(+4/–1)–C		
lacY	9–27	N–(+1/–0)		Buchel *et al.* (1980)
	47–65		(+1/–2)	
	77–95	(+3/–1)		
	103–121		(+0/–0)	
	144–162	(+5/–3)		
	168–186		(+0/–0)	
	351–369			
	383–401		(+0/–1)	
		(+2/–2)–C		
lep	4–22		N–(+0/–0)	Wolfe *et al.* (1983)
	58–76	(+10/–5)		
			[+25/–30]–C	

Table I cont.

B. Proteins with predicted topology				
hisQ	19–37		N–(+0/–1)	Higgins *et al.* (1982)
	62–80	(+3/–2)		
	94–112		(+0/–4)	
	190–208	[+8/–5]		
			(+4/–3)–C	
cds	20–38	N–(+2/–1)		Icho *et al.* (1985)
	51–69		(+1/–2)	
	85–103	(+3/–0)		
	117–135		(+1/–2)	
	155–173	(+5/–1)		
	179–197		(+0/–1)	
	228–246	(+4/–6)		
			(+1/–1)–C	
uncC	13–31		N–(+0/–2)	Kanazawa *et al.* (1982)
	53–71	(+3/–3)		
			(+0/–1)–C	
secY	23–41	N–(+5/–2)		Cerretti *et al.* (1983)
	77–95		(+4/–3)	
	122–140	(+6/–3)		
	154–172		(+0/+0)	
	186–204	(+1/–2)		
	217–235		(+1/–2)	
	277–295	(+8/–1)		
	319–337		(+0/–0)	
	373–391	(+7/–4)		
	399–417		(2+/–2)	
		(+4/–3)–C		
malG	19–37	N–(+3/–0)		Dassa and Hofnung (1985)
	92–110		(+4/–4)	
	126–144	(+4/–0)		
	159–177		(+1/–2)	
	206–224	(+2/–5)		
	263–281		(+1/–3)	
		(+2/–1)–C		
livH	21–39		N–(+0/–1)	Nazos *et al.* (1986)
	48–66	(+0/–1)		
	71–89		(+0/–1)	
	105–123	(+5/–1)		
	156–174		(+1/–3)	
	205–223	(+6/–3)		
	246–264		(+1/–0)	
	282–300	(+1/–3)		
			(2+/–3)–C	
uncB	46–64		N–(+1/–5)	Gay and Walker (1981)
	100–118	(+7/–2)		
	148–166		(+1/–4)	
	220–238	(+5/–3)		
	246–264		(+0/–0)	
		(+0/–3)–C		
unc C130	18–36	N–(+4/–0)		Gay and Walker (1981)
	45–63		(+1/–1)	
	79–97	(+4/–1)		
	103–121		(+1/–0)	
		(+1/–1)–C		
frd 13 kd protein	27–45	N–(+2/–2)		Grundström and Jaurin (1982)
	62–80		(+2/–2)	
	99–117	(+3/–1)		
			(+0/–1)–C	
pstA	35–53	N–(+8/–2)		Surin *et al.* (1985)
	88–106		(+1/–3)	
	128–146	(+3/–3)		
	152–170		(+0/–1)	
	201–219	(+6/–2)		
	268–286		(+2/–3)	
		(+3/–1)–C		

Table II. Amino acid frequencies (percent) in inner membrane proteins and in control samples of soluble cytosolic (s-cyt) and periplasmic (s-per) proteins

Residue	'Known' topology				'Predicted' topology					
	cyt	per	N_{in}	N_{out}	cyt	per	N_{in}	N_{out}	s-cyt	s-per
Ala	11.3	8.7	11.6	11.7	10.3	8.5	11.4	12.4	9.6	9.0
Cys	0.2	0.4	2.4	1.5	0.2	0.2	1.4	1.2	1.0	0.5
Asp	6.3	4.0	0.3	0.9	4.0	4.5	0.3	0.6	5.9	6.8
Glu	4.9	3.2	0.3	0.0	5.0	3.9	0.3	0.1	7.0	5.0
Phe	5.3	6.6	13.9	9.9	4.0	5.5	10.9	9.4	3.6	3.3
Gly	8.5	11.5	8.2	9.6	9.4	9.8	7.8	9.5	7.4	9.5
His	1.4	3.0	1.1	0.6	2.1	2.6	0.4	0.5	2.2	1.7
Ile	4.3	4.9	10.3	9.6	4.5	5.3	12.4	11.8	5.7	4.8
Lys	7.9	1.9	1.1	0.6	8.1	1.7	0.8	0.2	5.8	6.4
Leu	6.9	8.7	16.8	21.9	8.0	10.6	19.0	19.4	9.4	7.0
Met	2.4	2.1	5.0	3.5	3.4	3.2	4.9	3.9	2.5	1.6
Asn	2.8	5.3	0.3	0.9	3.3	5.2	0.5	0.7	3.8	6.0
Pro	3.8	8.5	3.7	0.9	4.1	7.3	3.9	1.4	4.1	4.2
Gln	3.2	4.0	0.3	1.2	3.4	4.6	0.9	1.2	4.2	4.2
Arg	7.9	2.3	0.3	0.0	8.0	2.9	0.3	0.2	6.5	3.6
Ser	5.9	4.9	5.5	4.4	6.0	5.4	5.2	4.3	5.3	6.7
Thr	4.0	5.3	4.2	5.6	4.7	5.3	4.1	5.1	5.0	6.9
Val	6.5	3.6	10.5	9.1	6.1	5.1	11.2	11.2	7.4	6.6
Trp	2.2	5.3	2.6	4.7	2.0	3.9	2.3	3.5	1.0	1.6
Tyr	4.3	5.7	1.8	3.5	3.5	4.4	2.0	3.4	2.6	4.5
Total	494	471	380	342	1206	1056	912	855	45 699	15 258

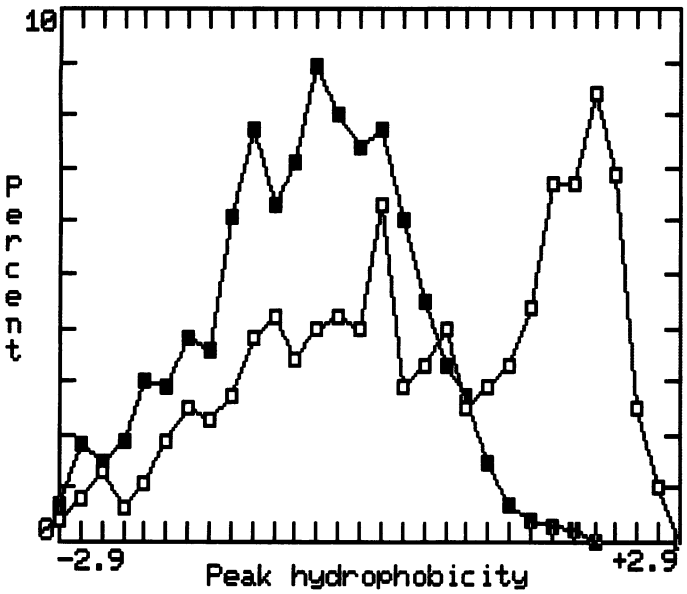

Fig. 1. Distribution of hydrophobicity peak heights calculated with a window of 19 residues, see materials and methods. Open squares: 66 bacterial inner membrane proteins (522 peaks in total); solid squares: 72 bacterial soluble cytosolic proteins (1495 peaks in total). Positive values are more hydrophobic.

peak hydrophobicity distribution for the inner membrane proteins (corresponding to connecting loops and spanning segments) stands out clearly. Due to the poor separation between the two peaks, however, segments with a peak hydrophobicity value in the range 0.8 – 1.4 cannot be unambiguously assigned to any of the two groups. Thus, from the initial 66 proteins, 10 which had *no* segment with a peak hydrophobicity in this critical range, together with the 10 well-characterized proteins discussed above, were selected as being reasonably likely to have correctly predicted spanning segments, also listed in Table I. For these proteins it was assumed that their trans-membrane topologies are such that a minimum number of positively charged residues are placed in the periplasmic loops, and the amino acid counts were again collected as above, Table II. Aside from the same 4-fold difference in the frequency of Arg + Lys as observed in the smaller sample, the incidence of Pro is still significantly higher ($P < 0.001$) in the periplasmic than in the cytosolic loops (f_{Pro} = 7.3% versus 4.1%), whereas there is no longer any difference in the frequency of Asp + Glu (8.4% versus 9.0%). Pro is also still significantly reduced in frequency ($P < 0.001$) only in the spanning segments with the N-terminus facing the periplasm and not in those with the opposite orientation (f_{Pro} = 1.4% versus 3.9%). No strong preferences for specific positions within the spanning segments were found for any of the residues. All these observations hold true even when the 10 'predicted' proteins are considered alone.

So far, only mean frequencies for the whole sample have been discussed; however, as shown in Figure 2, *all* periplasmic loops seem to be similarly reduced in their amount of positively charged residues. The distributions of the number of negatively charged residues, on the other hand, do not differ significantly between periplasmic and cytosolic loops (data not shown).

Long periplasmic loops have normal Arg and Lys frequencies

In the analysis above, only relatively short connecting loops (less than 65 residues long and with a mean length of ~20 residues) have been considered. Some inner membrane proteins have much longer periplasmic loops; thus *malF* has a periplasmic domain some 185 residues long, and the chemotaxis proteins *tar, tap, tsr*, (Krikos *et al.*, 1983) and *trg* (Bollinger *et al.*, 1984) have periplasmic domains counting about 165 residues. The overall amino acid compositions of these domains, however, are not significantly different from the samples of soluble cytosolic and periplasmic proteins, with $f_{Arg+Lys}$ = 10.4% and $f_{Asp+Glu}$ = 10.8% (data not shown).

The number of positively (but not negatively) charged residues tends to alternate between successive loops

If a local alternation in the number of positively charged residues between successive connecting loops is an important characteristic

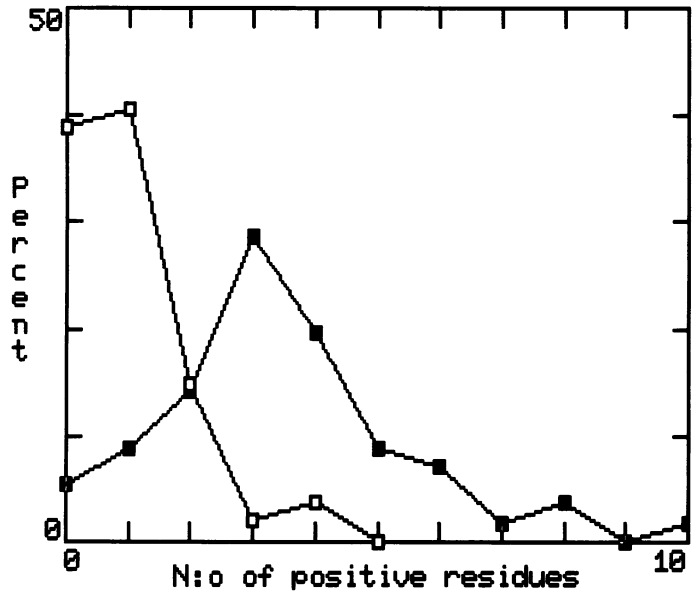

Fig. 2. Distribution of the number of positively charged residues in periplasmic connecting loops (open squares, 54 loops in total) and cytosolic connecting loops (solid squares, 56 loops in total) in the 20 inner membrane proteins listed in Table I.

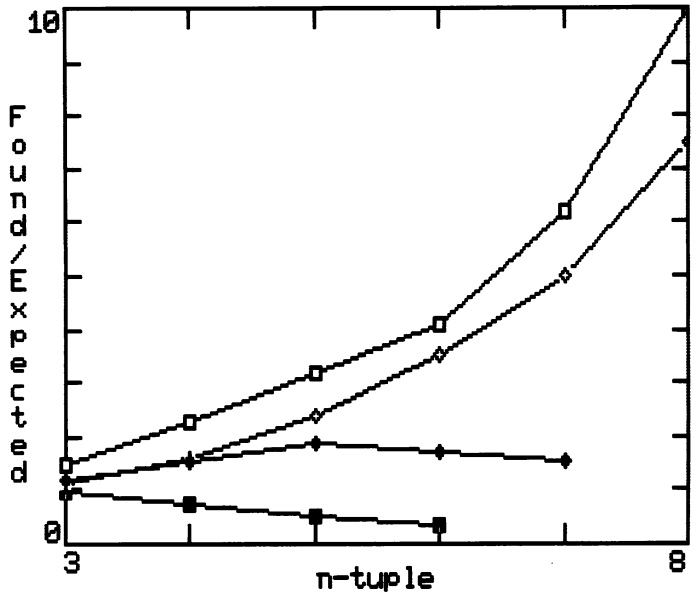

Fig. 3. Quotient between the number of strictly alternating *n*-tuples found in a sample of 66 inner membrane proteins and the number obtained for a sample consisting of 10 randomly scrambled copies of each of the original entries, see Materials and methods (open squares: positively charged residues; solid squares: negatively charged residues; open diamonds: total number of charged residues; solid diamonds: net charge). For the positively charged residues, the absolute numbers of strictly alternative *n*-tuples observed for the inner membrane protein sample are: 94 (63.9 expected) out of 121 3-tuples, 62 (27.4 expected) out of 94 4-tuples, 40 (12.6 expected) out of 71 5-tuples, 22 (5.4 expected) out of 53 6-tuples, 13 (2.1 expected) out of 38 7-tuples, and 8 (0.8 expected) out of 27 8-tuples.

of bacterial inner membrane proteins, then this should show up in an analysis of the pattern of alternation in subsequences including three, four, five, etc. connecting segments. Thus I have looked for the number of strictly alternating *n*-tuples (triplets, quadruplets, etc.; see Methods) in the series of numbers that one gets by noting the number of positively charged residues (or negatively charged residues, or the net charge, or the total charge) in successive connecting loops, both for the well-characterized proteins and from proteins with spanning segments predicted from hydrophobicity analysis. An occasional error in the prediction will have no great effect on the results in this case, and I have thus used the whole 66-protein sample and required a peak hydrophobicity >1.2 for predicting a spanning segment. As a further restriction, I have required that no connecting loop be longer than 65 residues (when a longer loop was predicted, the protein was divided into two independent sequences).

The results of this anlaysis are shown in Figure 3 where the quotient between the number of observed strictly alternating *n*-tuples and the mean number of such *n*-tuples in a sample consisting of 10 randomly scrambled copies of each of the original series is plotted against *n*; again the positively charged residues (followed by the total number of charged residues) stand out as the characteristic giving the largest number of strictly alternating *n*-tuples for all values of *n*.

A 'grammar' for membrane protein topology

Although the number of proteins with reasonably well-defined trans-membrane topology is limited, one can still find examples ranging from the simplest possible case — a single membrane spanning segment bounded by two short exposed regions — to rather complex cases such as *malF* which most likely has eight spanning segments and a large periplasmic domain between segments 3 and 4. A close inspection of Table I shows the *Rps. capsulata* LH1 and LH2 polypeptides behaving as expected if the number of positively charged residues determine the topology: the region with the higher number of Arg + Lys and the higher total charge faces the cytosol.

The phage M13 major coat protein precursor and *E. coli* leader pepidase (*lep*) are on the next level of complexity with two likely spanning segments (see Materials and methods). Again, the orientation is as expected from the distribution of the positively charged residues; note in particular that the first hydrophobic region in *lep* probably has its N-terminus facing the periplasm which correlates with a lack of basic residues in the extra-membraneous N-terminal region. The same situation is observed for the single spanning segment in the *Rps. viridis* reaction centre H-subunit.

Among the proteins with multiple spanning segments, the *Rps. viridis* reaction centre L subunit shows a perfect series of alternating numbers of Arg + Lys; the M-subunit, *malF*, and *lacY* also conform to the rule with the qualification that a couple of neighbouring connecting loops have the same positive charge; only bacteriorhodopsin shows one 'inversion' breaking the strict alternation. It thus appears that the topology of the bacterial inner membrane proteins of the 'complex' kind can be generated by applying the rules found to hold for the simple one- and two-spanning segment proteins.

Implications for membrane protein biogenesis and protein secretion in bacteria

The surprisingly good correlation between the distribution of positively charged residues and the trans-membrane topology of a relatively large number of functionally diverse proteins discussed above may lead one to think of the biogenesis of bacterial inner membrane proteins primarily in terms of the 'helical hairpin' hypothesis (see Introduction). Within the framework of this model, it is easy to see how membrane integration of the protein could be determined by two factors: the strengths of the hydrophobic interaction between the putative spanning segments and the membrane, and the activation energy barriers associated with the translocation of the periplasmic connecting loops through the membrane. Thus, in a post-translational 'helical hairpin'

mechanism, the most hydrophobic spanning segment may insert first together with the one of its nearest neighbours that allows formation of the helical hairpin with the most easily translocatable connecting loop, followed by insertion of less hydrophobic 'hairpins'. Single N- and C-terminal spanning segments such as the N-terminal segment in the *Rps. viridis* H-subunit or the C-terminal spanning segments in the L and M subunits may insert as unpaired segments thus bringing the polar (but not highly positively charged) terminal residues through the membrane.

Kinetically determined 'locally optimal' insertions of this kind may in principle lead to structures with unpaired internal hydrophobic segments left on the cytosolic side that it should be possible to create by gene fusion, thus providing a critical test of the model. At this point, the most encouraging experimental finding is perhaps the observation that phage M13 coat protein precursor inserts spontaneously in the correct orientation even into protein-free liposomes (Geller and Wickner, 1985).

Presumably a mechanism such as this can only work for relatively short connecting loops, no more than perhaps 60 – 70 residues long (i.e. the spanning segments must be able to 'drag' the whole connecting loop into the membrane). Unfortunately, there are not yet any good examples of periplasmic loops with a length in the critical range between 60 and 100 residues; shorter loops have very few positively charged residues as demonstrated above, whereas longer loops such as found in *malF* and the chemotaxis proteins (165 – 185 residues long) show no such deficiency. The exact point of transition between loops with low and normal Arg + Lys counts thus remains to be determined. It is tempting to speculate that longer periplasmic loops are translocated across the membrane in the same way as soluble periplasmic proteins, i.e. in some sort of energy-driven (Chen and Tai, 1985), possibly post-translational (Randall, 1983) process initiated either by a cleavable N-terminal signal peptide or by an unpaired spanning segment on the N-terminal side of the periplasmic loop.

As for the amino acid composition of the spanning segments, the observation that Pro seems to be much more easily accommodated in those segments that have their N-terminus pointing towards the cytosol (N_{in}) than in those in the opposite orientation (N_{out}) is hard to explain. It does, however, cast some doubt on the hypothesis that the relatively high proline content in spanning segments from transport proteins as opposed to non-transport proteins has something to do with proline *cis – trans* isomerization being important for their transport function (Brandl and Deber, 1986), since most of the non-transport proteins analysed in that study are 'simple' membrane proteins with only one spanning segment in the N_{out}-orientation. Thus the difference in proline content observed by these authors may be a result of constraints imposed on the biogenesis of N_{in} versus N_{out} spanning segments, rather than transport protein function *per se*.

Why, finally, are positively charged residues apparently more critical in the connecting loops than negatively charged ones? If an interaction with the membrane potential is the decisive factor in the insertion process one would expect the total net charge rather than the positive charge or the total charge to be most strongly correlated with the trans-membrane topology; this does not seem to be the case. An alternative, though not mutually exclusive explanation more in keeping with the amino acid composition data is that the dipolar nature of the membrane, which is independent of any imposed potential, is at the root of the matter: the dipoles associated with the lipid headgroups make the membrane more easily penetrated by anions than by cations, with an estimated difference in activation free energy of up to 10 kcal/mol (Flewelling and Hubbell, 1986). Thus kinetics, rather than equilibrium thermodynamics, may ultimately be determining the folding of bacterial inner membrane proteins.

Materials and methods

Hydrophobicity analysis

In the *Rps. viridis* reaction centre complex, all membrane spanning helices have at least 19 contiguous uncharged residues, and are at least 24 residues long (Michel *et al.*, 1986). Thus, hydrophobicity analysis was performed using a 19-residue moving window and the Engelman – Steitz hydrophobicity scale (Engelman and Steitz, 1981). To partition the sequence into non-overlapping membrane spanning segments and connecting loops, the highest peak in the hydrophobicity profile was located, and a segment of 23 residues (19 residues in the spanning segment proper and two additional residues added at both ends) was removed from further consideration. This procedure was repeated until no segment of length 23 residues or more remained. From the list of putative spanning segments thus obtained, those with a mean 19-residue hydrophobicity greater than a pre-set cutoff value (1.2 or 1.4 kcal/mol, see text) were predicted as true trans-membrane helices. When amino acid counts were taken, the two added amino acids at the ends of each 19-residue spanning segment were counted as belonging to the connecting loops.

Charge calculation

In all charge calculations, Arg and Lys were counted as +1, Asp and Glu as −1. C-terminal carboxyl groups were also counted as −1. N-terminal amino groups were not counted, since (i) these may be formylated and hence uncharged during biogenesis and membrane integration of the protein, and (ii) since this group has a significantly lower pK_a than the basic moieties on the Arg and Lys side chains (around 9.5 versus 12.5 and 10.5; Bohinski, 1973).

Sequence samples

Well-characterized inner membrane proteins. This group contained the whole or parts of 10 proteins:

Rps. viridis reaction centre L, M, and H subunits: the best characterized of all membrane proteins to date, with the full three-dimensional X-ray structure having been determined (Deisenhofer *et al.*, 1985).

H. halobium bacteriorhodopsin: this protein is known from electron microscopy to have seven membrane spanning segments (Henderson and Unwin, 1975). The membrane topology of the chain has also been well mapped by protease cleavage experiments (Engelman and Steitz, 1984).

Rps. capsulata light-harvesting complex LH1 and LH2 polypeptides: these short polypeptides (58 and 49 residues long) have only one membrane spanning segment. Their N- and C-termini have been mapped to the cytosolic and periplasmic side of the membrane, respectively (Tadros *et al.*, 1986).

E. coli malF: the topology of the first three membrane spanning segments and the following long periplasmic domain has been mapped by *lacZ* and Tn*phoA* transposon fusions (Manoil,C., Boyd,D. and Froshauer,S., personal communication; see also Froshauer and Beckwith, 1984; Manoil and Beckwith, 1985). The C-terminal membrane domain has been less well mapped, but the existence of five unambiguous (mean 19-residue hydrophobicity > 1.4) putative spanning segments in this domain makes the topology shown in Table I highly probable.

Phage M13 major coat protein: this is a short protein with an N-terminal cleavable signal peptide, a periplasmic loop, and a C-terminal spanning segment. It integrates into membranes as a 'helical hairpin' in the absence of leader peptidase (Geller and Wickner, 1985).

E. coli lactose permease (*lacY*): the N-terminus, the C-terminus, and an exposed segment around residue 135 have been mapped to the cytosolic side of the membrane (Bieseler *et al.*, 1985; Seckler *et al.*, 1983,1986). Six unambiguous putative spanning segments in the N-terminal half and two close to the C-terminus make it possible to derive the partial topology shown in Table I (see also Vogel *et al.*, 1985).

E. coli leader peptidase (*lep*): this protein has a large C-terminal periplasmic domain, can be cleaved by trypsin attacking from the cytosolic side of the membrane around residue 50, and does not have a cleavable signal peptide (Wolf *et al.*, 1983). The existence of two unambiguous putative spanning segments (residues 2 – 24 and 56 – 78) strongly suggests the topology given in Table I.

Inner membrane proteins with predicted topology

Fifty-six additional sequences of bacterial inner membrane proteins were extracted from the NBRF Protein Sequence Database (Release 7.0) or collected from the literature. Ten of these were found to have only unambiguous (mean 19-residue hydrophobicity > 1.4 kcal/mol) and no ambiguous (mean 19 residue hydrophobicity between 0.8 and 1.4 kcal/mol, see text) putative spanning segments with connecting loops shorter than 65 residues. Their topologies were predicted assuming that a minimum number of postively charged residues are placed in periplasmic loops, see Table I.

Cytosolic and periplasmic reference sets

72 sequences (45 699 residues) of soluble cytosolic and 45 sequences (15 258 residues) of soluble periplasmic or extracellular bacterial proteins were extracted from the NBRF Protein Sequence Database or collected from the literature. Signal peptides were removed from the latter group prior to analysis.

n-Tuple analysis

For each protein analysed, the number of positively charged residues in successive connecting loops was recorded as a series of numbers, e.g. 3,1,4,3,3,1. Then, the number of strictly alternating *n*-tuples in this series was noted, e.g. two 3-tuples (3,1,4; 1,4,3) and one 4-tuple (3,1,4,3) in this example. The total numbers of such *n*-tuples in the whole sample was compared with the numbers obtained when the original series were randomly scrambled 10 times.

Statistical analysis

The statistical significance of observed differences in amino acid composition was assessed using χ^2-analysis.

Acknowledgement

This work was supported by a grant from the Swedish Natural Sciences Research Council.

References

Bieseler,B., Prinz,H. and Beyreuther,K. (1985) *Ann. N.Y. Acad. Sci.*, **456**, 309–325.

Blobel,G. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 1496–1500.

Bohinski,R.C. (1973) *Modern Concepts in Biochemistry*. Allen & Bacon, Boston, MA.

Bollinger,J., Park,C., Harayama,S. and Hazelbauer,G.L. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 3287–3291.

Brandl,C.J. and Deber,C.M. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 917–921.

Büchel,D.E., Gronenborn,B. and Müller-Hill,B. (1980) *Nature*, **283**, 541–545.

Cerretti,D.P., Dean,D., Davis,G.R., Bedwell,D.M. and Nomura,M. (1983) *Nucl. Acids Res.*, **11**, 2599–2616.

Chen,L. and Tai,P.C. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4384–4388.

Dassa,E. and Hofnung,M. (1985) *EMBO J.*, **4**, 2287–2293.

Deisenhofer,J., Epp,O., Miki,K., Huber,R. and Michel,H. (1985) *Nature*, **318**, 618–624.

Drews,G. (1985) *Microbiol. Rev.*, **49**, 59–70.

Dunn,R., McCoy,J., Simsek,M., Majumdar,A., Chang,S.H., RajBhandary,U.L. and Khorana,H.G. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 6744–6748.

Engelman,D.M. and Steitz,T.A. (1981) *Cell*, **23**, 411–422.

Engelman,D.M. and Steitz,T.A. (1984) In Wetlaufer (ed.), *The Protein Folding Problem*. AAAS, New York, pp. 87–113.

Flewelling,R.F. and Hubbell,W.L. (1986) *Biophys. J.*, **49**, 541–552.

Froshauer,S. and Beckwith,J. (1984) *J. Biol. Chem.*, **259**, 10896–10903.

Gay,N.J. and Walker,J.E. (1981) *Nucl. Acids Res.*, **9**, 3919–3926.

Geller,B.L. and Wickner,W. (1985) *J. Biol. Chem.*, **260**, 13281–13285.

Grundström,T. and Jaurin,B. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 1111–1115.

Henderson,R. and Unwin,P.N.T. (1975) *Nature*, **257**, 28–32.

Higgins,C.F., Haag,P.D., Nikaido,K., Ardeshir,F., Garcia,G. and Ames,G.F.L. (1982) *Nature*, **298**, 723–727.

Icho,T., Sparrow,C.P. and Raetz,C.R.H. (1985) *J. Biol. Chem.*, **260**, 12078–12083.

Kanazawa,H., Kayano,T., Kiyasu,T. and Futai,M. (1982) *Biochem. Biophys. Res. Commun.*, **105**, 1257–1264.

Krikos,A., Mutoh,N., Boyd,A. and Simon,M.I. (1983) *Cell*, **33**, 615–622.

Manoil,C. and Beckwith,J. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 8129–8133.

Michel,H., Weyer,K.A., Gruenberg,H. and Lottspeich,F. (1985) *EMBO J.*, **4**, 1667–1672.

Michel,H., Weyer,K.A., Gruenberg,H., Dunger,I., Oesterhelt,D. and Lottspeich,F. (1986) *EMBO J.*, **5**, 1149–1158.

Nazos,P.M., Antonucci,T.K., Landick,R. and Oxender,D.L. (1986) *J. Bacteriol.*, **166**, 565–573.

Randall,L.L. (1983) *Cell*, **33**, 231–240.

Seckler,R., Wright,J.K. and Overath,P. (1983) *J. Biol. Chem.*, **258**, 10817–10820.

Seckler,R., Möröy,T., Wright,J.K. and Overath,P. (1986) *Biochemistry*, **25**, 2403–2409.

Surin,B.P., Rosenberg,H. and Cox,G.B. (1985) *J. Bacteriol.*, **161**, 189–198.

Tadros,M.H., Frank,R. and Drews,G. (1986) *FEBS Lett.*, **196**, 233–236.

van Wezenbeek,P.M.G.F., Hulsebos,T.J.M. and Schoenmakers,J.G.G. (1980) *Gene*, **11**, 129–148.

Vogel,H., Wright,J.K. and Jähnig,F. (1985) *EMBO J.*, **4**, 3625–3631.

von Heijne,G. (1985) *Current Topics in Membranes and Transport*, **24**, 151–179.

Wolfe,P.B., Wickner,W. and Goodman,J.M. (1983) *J. Biol. Chem.*, **258**, 12073–12080.

Received on 21 July 1986