ARTICLES

# Molecular code for transmembrane-helix recognition by the Sec61 translocon

Tara Hessa[1]*, Nadja M. Meindl-Beinker[1]*, Andreas Bernsel[2]*, Hyun Kim[1], Yoko Sato[1], Mirjam Lerch-Bader[1], IngMarie Nilsson[1], Stephen H. White[3] & Gunnar von Heijne[1,2]

**Transmembrane α-helices in integral membrane proteins are recognized co-translationally and inserted into the membrane of the endoplasmic reticulum by the Sec61 translocon. A full quantitative description of this phenomenon, linking amino acid sequence to membrane insertion efficiency, is still lacking. Here, using *in vitro* translation of a model protein in the presence of dog pancreas rough microsomes to analyse a large number of systematically designed hydrophobic segments, we present a quantitative analysis of the position-dependent contribution of all 20 amino acids to membrane insertion efficiency, as well as of the effects of transmembrane segment length and flanking amino acids. The emerging picture of translocon-mediated transmembrane helix assembly is simple, with the critical sequence characteristics mirroring the physical properties of the lipid bilayer.**

Most integral membrane proteins are composed of bundles of tightly packed transmembrane (TM) α-helices[1]. The lipid bilayers into which these proteins are inserted are highly anisotropic and their physicochemical characteristics vary markedly over short distances[2]. This anisotropy is reflected in the distribution of different amino acids in the membrane-embedded parts of integral membrane proteins[3], but the actual recognition of TM helices in nascent polypeptide chains is performed by so-called translocons, complex molecular machines that ensure both the translocation of globular proteins across membranes and the integration of membrane proteins into membranes[4].

What is the 'molecular code' that allows a translocon to recognize TM helices in newly synthesized membrane proteins? We have recently described a system in which $\Delta G_{app}$, the apparent free energy of insertion of a TM helix into the membrane of the endoplasmic reticulum, can be measured and have presented a 'biological' hydrophobicity scale based on such measurements[5]. This scale quantifies the contribution of each of the 20 amino acids, $\Delta G^{aa}_{app}$, to $\Delta G_{app}$ for residues placed in the middle position in a 19-residue TM helix. However, if, as has been suggested[5,6], the recognition of TM helices by the Sec61 translocon is based on a thermodynamic partitioning into the anisotropic environment of the lipid bilayer, $\Delta G^{aa}_{app}$ should vary with the residue's position in the membrane[7]. $\Delta G_{app}$ may also be expected to vary with the overall length of the TM helix, and possibly with the nature of the residues immediately flanking the helix.

To arrive at a full quantitative description of TM helix recognition by the Sec61 translocon, we used the experimental setup summarized in Fig. 1. In brief, systematically designed test segments (H-segments) are introduced near the middle of the large luminal P2 domain of the model protein Lep; the protein is then expressed *in vitro* in the presence of endoplasmic-reticulum-derived dog pancreas rough microsomes, and an apparent equilibrium constant for membrane integration of the H-segment is calculated on the basis of the amount of singly versus doubly glycosylated protein[5]. This in turn can be converted into an apparent free energy of membrane insertion, $\Delta G_{app}$ (see Methods). Control experiments show that the identity of the TM1 and TM2 helices in Lep has little influence on $\Delta G_{app}$ (ref. 8).

## Position-dependent contributions to $\Delta G_{app}$

To obtain a comprehensive data set describing the positional variability in $\Delta G^{aa}_{app}$ for the 20 amino acids, we designed a set of Lep constructs in which each kind of residue was systematically scanned across a Leu-Ala-based H-segment and $\Delta G_{app}$ values were measured. All H-segments were designed with the sequence GGPG-(19 residues)-GPGG. For each residue type, the numbers of Leu and Ala residues in the H-segment were chosen such that $\Delta G_{app} \approx 0$ kcal mol$^{-1}$ (1 kcal ≈ 4.18 kJ) when the residue was in the middle position of the 19-residue stretch. The results show that $\Delta G^{aa}_{app}$ values vary strongly with position for charged and highly polar

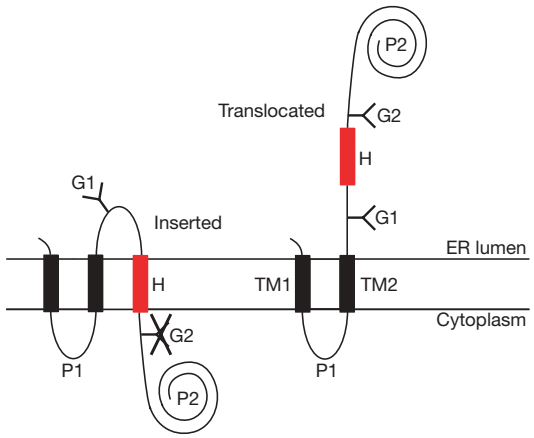


**Figure 1 | The Lep model protein.** *Escherichia coli* leader peptidase (Lep) has two TM helices (TM1 and TM2) and a large luminal domain (P2). It inserts into rough microsomes in an $N_{lum}$–$C_{lum}$ orientation. H-segments (red) are engineered into the P2 domain with two flanking Asn-X-Thr glycosylation acceptor sites (G1, G2). Constructs for which the H-segment is integrated into the endoplasmic reticulum membrane as a TM helix are glycosylated only on the G1 site (left), whereas those for which the H-segment is translocated across the membrane are glycosylated on both the G1 and G2 sites (right).

[1]Center for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden. [2]Stockholm Bioinformatics Center, AlbaNova, Stockholm University, SE-106 91 Stockholm, Sweden. [3]Department of Physiology and Biophysics and the Center for Biomembrane Systems, University of California at Irvine, Irvine, California 92697-4560, USA.
*These authors contributed equally to this work.

residues as well as for Pro (a strong helix-breaker), whereas they are nearly independent of position for weakly polar and apolar residues (Supplementary Fig. 1a). Although single charged residues may cause a shift in the position of the H-segment relative to the membrane and lead to an underestimate of $\Delta G^{aa}_{app}$, previous work suggests that such shifts are small (not more than 3 Å) and restricted to positions near the ends of the 19-residue H-segments used here[9,10].

Although the single-residue scans give a good first impression of the positional dependence of the $\Delta G^{aa}_{app}$ values, we sought to incorporate as much information as possible, both from the H-segments analysed previously[5,7] and from the constructs made in this study (Supplementary Table 1), to derive an optimized matrix ($\overline{\Delta G^{aa}_{app}}$) of position-specific $\Delta G^{aa}_{app}$ values. To this end, 324 19-residue H-segments for which the measured $\Delta G_{app}$ values are between $-1.5$ and $+1.0$ kcal mol$^{-1}$ (that is, within an interval where the accuracy in the $\Delta G_{app}$ determination is good) were collected. Using this data set, we performed a least-squares optimization in which the $\overline{\Delta G^{aa}_{app}}$ matrix elements for each residue were described by a gaussian function (see Methods). We also included a contribution from the hydrophobic moment of each H-segment[5]. Except for the hydrophobic moment part, H-segment $\Delta G_{app}$ values were modelled as a linear sum of free-energy values for individual amino acids:

$$\Delta G^{pred}_{app} = \sum_{i=1}^{l} \Delta G^{aa(i)}_{app} + c_0\mu \qquad (1)$$

where $l$ is the length of the segment (here, $l = 19$), $\Delta G^{aa(i)}_{app}$ is the matrix element giving the contribution from amino acid aa in position $i$, $\mu$ is the hydrophobic moment (see Methods), and $c_0$ is the weight parameter for the hydrophobic moment. The optimized $\overline{\Delta G^{aa}_{app}}$ matrix (Supplementary Table 2) was derived by minimizing the sum of the squared differences between the predicted $\Delta G_{app}$ values ($\Delta G^{pred}_{app}$) and measured $\Delta G_{app}$ values.

As expected, equation (1) reproduces the experimental single-residue scans well (Supplementary Fig. 1a) and also reproduces data from symmetrical pair-scans[5] in which two residues were scanned symmetrically from the centre of the H-segment to preclude shifts in the location of the H-segment relative to the membrane (Supplementary Fig. 1b). There is only one residue, proline, for which the single-scan and pair-scan results are qualitatively different: Pro has a fairly symmetric single-scan profile, but two Pro residues placed near each other in the centre of the H-segment are tolerated better than when they are spaced farther apart. This cooperative effect cannot be captured by the simple additive model in equation (1).

The optimized position-dependent gaussians describing the $\overline{\Delta G^{aa}_{app}}$ matrix are shown in Fig. 2 (blue curves). Figure 2 also shows statistical free-energy profiles derived from the distribution of the different amino acids in high-resolution membrane protein three-dimensional structures (red curves; see Methods); these profiles presumably reflect mainly interaction free energies between amino-acid side chains and the lipid bilayer. In general, the two sets of profiles match each other well. The profiles for His match rather poorly, however. A possible explanation is that a number of the known three-dimensional structures contain cofactor-binding His residues. Indeed, the statistical His profile obtained when all such

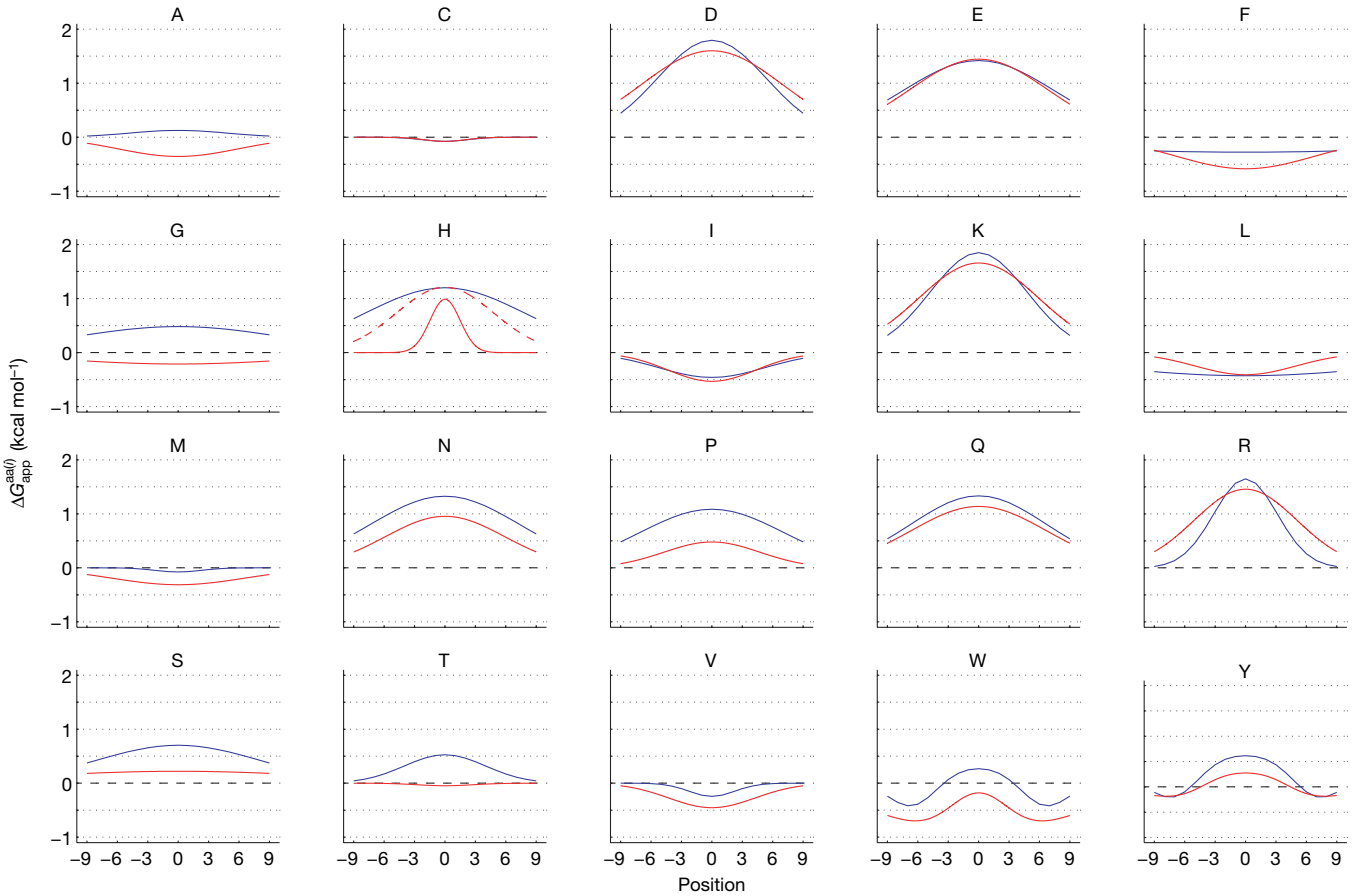


**Figure 2 | Position-specific $\Delta G_{app}$ contributions.** The gaussians describing the $\overline{\Delta G^{aa}_{app}}$ matrix—that is, the contribution from each individual amino acid to $\Delta G_{app}$—are plotted as a function of position within the 19-residue segment (blue). Amino acids are identified by their one-letter abbreviations. Position-specific statistical distributions calculated from three-dimensional structures of membrane proteins are shown in red. The dashed red line for His shows the statistical distribution obtained when all cofactor-containing proteins are omitted. To compare the profiles, one amino acid was equated to a $z$-coordinate displacement of 1.5 Å.

proteins are omitted matches the experimental profile much better (dashed red line).

How well can equation (1) predict $\Delta G_{app}$ values for H-segments in the training set or chosen from natural proteins? For 90% of the H-segments in the training set, $\Delta G_{app}^{pred}$ values are within $\pm 0.45$ kcal mol$^{-1}$ of the measured $\Delta G_{app}$ values (Supplementary Fig. 2). Overall, $\Delta G_{app}$ is well predicted by equation (1) also for 16 representative 19-residue segments from membrane proteins of known structure (see below) and six additional segments that include sequences from a weakly hydrophobic single-span TM protein and five non-membrane proteins (Supplementary Fig. 2). The only outliers are a couple of very highly charged S4 helices from ion-channel voltage-sensor domains[11], for which equation (1) overestimates the cost of membrane insertion. We conclude that the simple additive model provides a good first approximation to $\Delta G_{app}$ for most natural protein sequences, unless they are exceptionally rich in charged residues or contain multiple proline residues.

### Relation between H-segment length and $\Delta G_{app}$

To delineate the relation between H-segment length and $\Delta G_{app}$, we analysed a series of Leu-Ala based constructs with the overall composition GGPG-(*n*L, *m*A)-GPGG, where $n = 0$, 1, 2, 3, 5, 7. The *m* values were chosen such that we could identify by interpolation, for each *n*, the *m* value for which the H-segment inserts into the membrane in 50% of the molecules ($m_{50}$, corresponding to $\Delta G_{app} = 0$ kcal mol$^{-1}$). In addition, we included one set of constructs with the overall composition GGPG-(*n*L)-GPGG. The results are shown in Fig. 3. As the number of Leu residues (*n*) decreases, the number of Ala residues (*m*) required to reach a given $\Delta G_{app}$ increases; in fact, as shown in the inset to Fig. 3, there is a striking linear correlation between the number of Leu and Ala residues in the H-segment required for $\Delta G_{app} = 0$ kcal mol$^{-1}$. The least-squares fit to the data points in Fig. 3 (inset) is $m_{50} = -2.9n + 26$; that is, for each Leu residue removed from the H-segment, 2.9 Ala have to be added to maintain $\Delta G_{app} = 0$ kcal mol$^{-1}$. The correlation holds over an extended range of H-segment lengths ($9 \leq n + m \leq 30$). Because the Leu side chain has a roughly 2.4-fold larger accessible surface area than the Ala side chain[12] (95 Å$^2$ compared with 40 Å$^2$), the Leu-Ala

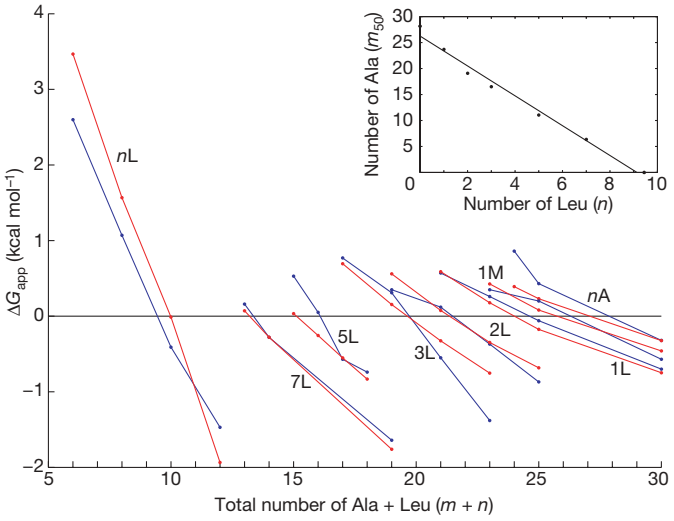


**Figure 3 | Length dependence of $\Delta G_{app}$.** Measured $\Delta G_{app}$ (blue) and predicted, cross-validated $\Delta G_{app}^{pred}$ values (red) for H-segments with the composition GGPG-(*n*L, *m*A)-GPGG and GGPG-(1M, *m*A)-GPGG. The lines connect data points with fixed *n* and varying *m*. The inset plots the number of Ala residues against the number of Leu residues required for $\Delta G_{app} = 0$ kcal mol$^{-1}$ obtained from the data in the main panel ($m_{50} = -2.9n + 26$; $R^2 = 0.99$). The data for the (1M, *n*A) constructs were not used in the optimization of the length-corrected equation (1).

based H-segments with $\Delta G_{app} = 0$ kcal mol$^{-1}$ have a roughly constant non-polar surface area of about 1,000 Å$^2$.

A second notable feature in Fig. 3 is that the slope of the lines for different *n* tends towards zero as the overall length $l = n + m$ increases. Closer inspection reveals that the derivative $\partial \Delta G_{app}/\partial l$ is roughly proportional to $l$ (Supplementary Fig. 3). This suggests that a phenomenological, length-dependent expression for $\Delta G_{app}^{pred}$ can be obtained from equation (1) (which holds for $l = 19$) to which is added an expression of the form $c_1 + c_2 l + c_3 l^2$; the best fit obtained by optimizing $c_1$, $c_2$ and $c_3$ (see Methods) is shown as red lines in Fig. 3.

To map the variation in $\Delta G_{app}^{aa(i)}$ values as a function of H-segment length, we measured position-dependent $\Delta G_{app}$ values for a single Lys residue in H-segments of different lengths (15, 19 and 25 residues; Supplementary Fig. 4a). A lengthening of the H-segment essentially results in a 'stretching' of the Lys profile, while maintaining the difference in $\Delta G_{app}$ between the middle and terminal positions at about 1.8 kcal mol$^{-1}$. We further tested whether the $\Delta G_{app}^{aa(i)}$ values for residues spaced at either end of a long H-segment are additive or whether there is a maximum length over which two residues cannot simultaneously contribute to $\Delta G_{app}$. To this end, we scanned two Leu residues symmetrically from the centre of a 25-residue H-segment. As found for a 19-residue H-segment[5], $\Delta G_{app}$ is roughly constant regardless of the spacing between the two Leu residues (Supplementary Fig. 4b). We also found that the contributions to $\Delta G_{app}$ from Trp residues introduced near the ends of the H-segment are roughly additive for both 19-residue and 25-residue H-segments (constructs 40, 363 and 417–422 in Supplementary Table 1). These results imply that even very long H-segments behave as one unit in terms of membrane insertion and that equation (1), corrected for length dependence and with the $\overline{\Delta G_{app}^{aa}}$ matrix suitably 'stretched' or 'compressed' for different lengths (see Methods), should provide a good model for TM helix recognition by the Sec61 translocon. A web server implementing the length-corrected model for $\Delta G_{app}^{pred}$ is available at http://www.cbr.su.se/DGpred/.

### Contributions to $\Delta G_{app}$ from flanking residues

The cytoplasmic ends of TM helices are often flanked by positively charged Lys and Arg residues[13], suggesting that flanking charged residues might contribute to $\Delta G_{app}$. We therefore tested two series of H-segments in which the central 19-residue stretch had the composition 3L/16A or 4L/15A. The Gly residues in the GGPG…GPGG flanks used above were replaced with Asp, Glu, Asn, Gln, Lys, Arg or Ser, and the luminal and cytoplasmic flanks were combined in different ways.

The changes in $\Delta G_{app}$ relative to the 3L/16A and 4L/15A H-segments with GGPG…GPGG flanks are shown in Supplementary Fig. 5a. The effects of Asp and Glu, as well as those of Asn and Gln, are strikingly different from the effects of Lys and Arg. Three Asp/Glu or Asn/Gln residues increase $\Delta G_{app}$ by about 0.9 kcal mol$^{-1}$ and about 0.5 kcal mol$^{-1}$, respectively, when present at the luminal end of the H-segment but not at its cytoplasmic end. In contrast, three Lys or Arg residues reduce $\Delta G_{app}$, both by $-0.7$ kcal mol$^{-1}$, when present at the cytoplasmic end but not at the luminal end. Ser has no appreciable effect on $\Delta G_{app}$ compared with Gly. The contributions to $\Delta G_{app}$ from flanking Asp and Lys residues are approximately additive, such that $\Delta\Delta G_{app}$ for H-segments with DDPD…KPKK flanks is close to that expected from adding the individual contributions (expected average $\Delta\Delta G_{app} = (0.8 - 0.7) = 0.1$ kcal mol$^{-1}$; observed average $\Delta\Delta G_{app} = 0.2$ kcal mol$^{-1}$). Additivity implies that the entire H-segment, including the flanking residues, may be recognized as one unit during membrane insertion.

To check whether this conclusion is valid also for other H-segment lengths, we made constructs with the same combinations of flanking residues for a 10-residue (10L) and a 25-residue (2L/23A) H-segment. As is clear from Supplementary Fig. 5b, the contributions from the charged flanking residues are additive even for the 25-residue

H-segment. Thus, flanking residues separated by as few as 10 and as many as 25 apolar residues (that is, spaced 15–38 Å apart) affect Sec61-mediated membrane integration of TM helices in the same way.

### $\Delta G^{pred}_{app}$ for TM helices in natural proteins

Finally, how well does the length-corrected equation (1) serve to identify TM helices in natural proteins? To address this question, we collected four test sets: first, all mammalian proteins annotated in SwissProt[14] as having a cleaved signal peptide and one single TM helix (349 single-spanning membrane proteins); second, all mammalian proteins annotated as having a cleaved signal peptide and no TM helix (1,012 soluble proteins targeted to the secretory pathway); third, all mammalian proteins annotated as located in the cytoplasm (670 cytoplasmic proteins); and fourth, all helix-bundle membrane proteins of known three-dimensional structure with at least two TM helices (508 TM helices from 66 Protein Data Bank[15] structures). The first, second and fourth sets contain proteins that have all passed through the endoplasmic reticulum translocon (or its prokaryotic SecYEG homologue), whereas the proteins in the third set have not visited the translocon. For the first to third sets, we identified in each protein (after removing the signal peptide) the segment with the lowest $\Delta G^{pred}_{app}$ value (for $17 \leq l \leq 33$), whereas for the fourth set we identified the segment with the lowest $\Delta G^{pred}_{app}$ value within each annotated TM helix (extended by ten residues at both the amino-terminal end and the carboxy-terminal end).

The results are summarized in Fig. 4. The overlap between the $\Delta G^{pred}_{app}$ distributions for the single-spanning transmembrane proteins and the secreted proteins is small, and the two distributions cross close to the zero-point on the scale defined by the experimental analysis of the designed H-segments. The discrimination between the two data sets is considerably better with the use of the $\Delta G^{pred}_{app}$ values than when simpler hydrophobicity scales are used (Supplementary Fig. 6).

The $\Delta G^{pred}_{app}$ distribution for the cytoplasmic proteins overlaps for the most part with that for the secreted proteins, as expected. There is, however, a significant number of cytoplasmic proteins with $\Delta G^{pred}_{app} < 0\,\mathrm{kcal\,mol^{-1}}$, as though the requirement to pass through the translocon has 'filtered out' proteins with such segments from the group of secreted proteins.

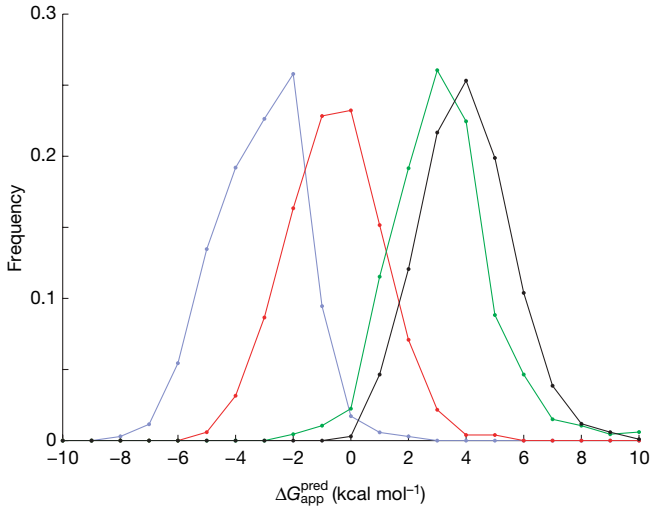


**Figure 4 | Distributions of $\Delta G^{pred}_{app}$ values in natural proteins.** The segment with lowest $\Delta G^{pred}_{app}$ was identified in 670 cytoplasmic (green), 1,012 secreted (black) and 349 single-spanning transmembrane proteins (blue); signal peptides were excluded. For 508 TM helices from multispanning membrane proteins of known three-dimensional structure (red), the segment with the lowest $\Delta G^{pred}_{app}$ for each helix was identified. Dots show the relative frequency of proteins with $\Delta G^{pred}_{app}$ within $\pm 0.5\,\mathrm{kcal\,mol^{-1}}$ of the value given on the *x* axis. None of the 1,012 secreted proteins has a segment with $\Delta G^{pred}_{app} < 0$, and only 3 of 349 transmembrane segments in the single-spanning proteins have $\Delta G^{pred}_{app} > 0$.

Although more data will be required for proper modelling of the quantitative effects on $\Delta G^{pred}_{app}$ of charged flanking residues, a rough estimate is that, on average, cytoplasmic positively charged flanking residues may decrease $\Delta G^{pred}_{app}$ by about $0.5\,\mathrm{kcal\,mol^{-1}}$ (M.L.-B., C. Lundin, H.K., I.N. and G.v.H., unpublished observations). Even with this correction, however, there is a surprisingly large fraction (about 25%) of the TM helices in the multi-spanning membrane proteins of known three-dimensional structure that have $\Delta G^{pred}_{app} > 0\,\mathrm{kcal\,mol^{-1}}$. Such segments would presumably be only inefficiently recognized as TM helices by the translocon if they were the only hydrophobic segment in a protein (as seen for the few that were tested in Supplementary Fig. 2). This suggests that a relatively large fraction of the TM helices in multi-spanning membrane proteins may depend on interactions with neighbouring TM helices for proper partitioning into the membrane. Indeed, several such cases have been described in the literature[16,17].

Our results show that the experimentally derived position-dependent $\Delta G^{aa(i)}_{app}$ profiles are similar to statistical residue-distribution profiles derived from TM helices in natural membrane proteins of known structure. $\Delta G^{pred}_{app}$ values obtained from a simple additive model, equation (1), are reasonably close to the $\Delta G_{app}$ values measured for H-segments of mixed amino-acid composition extracted from natural proteins, and they provide a good discrimination between TM helices in single-spanning mammalian membrane proteins and the most hydrophobic segments in mammalian secreted proteins. The relation between length and hydrophobicity for membrane insertion of Ala/Leu-based H-segments is a strikingly simple one, and H-segments as long as 25 residues behave as a single unit during membrane insertion; the simplest interpretation is that long H-segments can tilt or flex and thereby interact in their entirety with the lipid bilayer despite considerable 'hydrophobic mismatch'[18–21] and that the length-dependent contribution to $\Delta G_{app}$ approximates the free-energy cost associated with positive and negative mismatch between helix length and bilayer thickness. These results further support the idea that the recognition of TM helices by the Sec61 translocon critically involves a partitioning of the nascent polypeptide into the lipid bilayer[5,22], and they provide a quantitative basis for future studies of membrane protein biogenesis and prediction of membrane protein topology and structure.

## METHODS SUMMARY

Lep constructs were transcribed and translated in the TNT Quick coupled transcription–translation system supplemented with dog pancreas rough microsomes. The degree of membrane integration of each H-segment was quantified from SDS–PAGE gels by calculating an apparent equilibrium constant between the membrane-integrated and non-integrated forms: $K_{app} = f_{1g}/f_{2g}$, where $f_{1g}$ is the fraction of singly glycosylated Lep molecules and $f_{2g}$ is the fraction of doubly glycosylated Lep molecules, after correcting for the fact that a fully translocated P2 domain is glycosylated only to about 85% (ref. 5). The results were then converted to apparent free energies, $\Delta G_{app} = -RT\ln K_{app}$.

The full expression for the length-corrected equation (1) is

$$\Delta G^{pred}_{app} = \sum_{i=1}^{l} \Delta G^{aa(i)}_{app} + c_0 \sqrt{\left(\sum_{i=1}^{l} \Delta G^{aa(i)}_{app} \sin(100^\circ i)\right)^2 + \left(\sum_{i=1}^{l} \Delta G^{aa(i)}_{app} \cos(100^\circ i)\right)^2} + c_1 + c_2 l + c_3 l^2$$

with the optimized parameter values given in Supplementary Table 2.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 9 August; accepted 17 October 2007.**

1. Oberai, A., Ihm, Y., Kim, S. & Bowie, J. U. A limited universe of membrane protein families and folds. *Protein Sci.* **15**, 1723–1734 (2006).
2. Wiener, M. C. & White, S. H. Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of x-ray and neutron diffraction data. III. Complete structure. *Biophys. J.* **61**, 437–447 (1992).
3. Ulmschneider, M. B., Sansom, M. S. & Di Nola, A. Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins* **59**, 252–265 (2005).

4. Schnell, D. J. & Hebert, D. N. Protein translocons: multifunctional mediators of protein translocation across membranes. *Cell* **112,** 491–505 (2003).
5. Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433,** 377–381 (2005).
6. Heinrich, S., Mothes, W., Brunner, J. & Rapoport, T. The Sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. *Cell* **102,** 233–244 (2000).
7. Hessa, T., White, S. H. & von Heijne, G. Membrane insertion of a potassium channel voltage sensor. *Science* **307,** 1427 (2005).
8. Meindl-Beinker, N. M., Lundin, C., Nilsson, I., White, S. H. & von Heijne, G. Asn- and Asp-mediated interactions between transmembrane helices during translocon-mediated membrane protein assembly. *EMBO Rep.* **7,** 1111–1116 (2006).
9. Nilsson, I. *et al.* Proline-induced disruption of a transmembrane α-helix in its natural environment. *J. Mol. Biol.* **284,** 1165–1175 (1998).
10. Monné, M., Nilsson, I., Johansson, M., Elmhed, N. & von Heijne, G. Positively and negatively charged residues have different effects on the position in the membrane of a model transmembrane helix. *J. Mol. Biol.* **284,** 1177–1183 (1998).
11. Zhang, L. *et al.* Membrane insertion of the Shaker voltage sensor occurs both cotranslationally and posttranslationally. *Proc. Natl Acad. Sci. USA* **104,** 8263–8268 (2007).
12. Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105,** 1–12 (1976).
13. von Heijne, G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* **5,** 3021–3027 (1986).
14. O'Donovan, C. *et al.* High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **3,** 275–284 (2002).
15. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
16. Sadlish, H. & Skach, W. R. Biogenesis of CFTR and other polytopic membrane proteins: new roles for the ribosome–translocon complex. *J. Membr. Biol.* **202,** 115–126 (2004).
17. Buck, T. M., Wagner, J., Grund, S. & Skach, W. R. A novel tripartite motif involved in aquaporin topogenesis, monomer folding and tetramerization. *Nature Struct. Mol. Biol.* **14,** 762–769 (2007).
18. Killian, J. A. & von Heijne, G. How proteins adapt to a membrane–water interface. *Trends Biochem. Sci.* **25,** 429–434 (2000).
19. de Planque, M. R. R. & Killian, J. A. Protein–lipid interactions studied with designed transmembrane peptides: role of hydrophobic matching and interfacial anchoring. *Mol. Membr. Biol.* **20,** 271–284 (2003).
20. Yeagle, P. L., Bennett, M., Lemaitre, V. & Watts, A. Transmembrane helices of membrane proteins may flex to satisfy hydrophobic mismatch. *Biochim. Biophys. Acta* **1768,** 530–537 (2007).
21. Monné, M. & von Heijne, G. Effects of 'hydrophobic mismatch' on the location of transmembrane helices in the ER membrane. *FEBS Lett.* **496,** 96–100 (2001).
22. Heinrich, S. U. & Rapoport, T. A. Cooperation of transmembrane segments during the integration of a double-spanning protein into the ER membrane. *EMBO J.* **22,** 3654–3663 (2003).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank E. Missioux for technical assistance, and A. Elofsson and E. Lindahl for discussions. This work was supported by grants from the Swedish Foundation for Strategic Research, the Marianne and Marcus Wallenberg Foundation, the Swedish Cancer Foundation, the Swedish Research Council and the European Commission (BioSapiens) to G.v.H., the Magnus Bergvall Foundation to I.N., the National Institute of General Medical Sciences to S.H.W., the Swiss National Science Foundation to M.L.-B., and the Japan Society for the Promotion of Science to Y.S.

**Author Contributions** T.H. and N.M.M.-B. performed the experimental work together with H.K., Y.S., M.L.-B. and I.N. A.B. performed the computational work. T.H., N.M.M.-B., A.B., S.H.W. and G.v.H. prepared the manuscript. All authors discussed the results and commented on the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G.v.H. (gunnar@dbb.su.se).

## METHODS

**Enzymes and chemicals.** All enzymes, plasmid pGEM1, and the TNT Quick transcription–translation system were from New England Biolabs or Promega. [$^{35}$S]Met, deoxynucleotides and dideoxyribonucleotides were from GE Healthcare. Oligonucleotides were from Cybergene and MWG Biotech.

**Expression *in vitro* and quantification of membrane insertion efficiency.** All plasmids were constructed as described[23]. Constructs cloned in pGEM1 were transcribed and translated in the TNT Quick coupled transcription–translation system. The reaction was started by the addition of 1 μg of DNA template, 1 μl of [$^{35}$S]Met (15 μCi), and 1 μl of dog pancreas rough microsomes (a gift from M. Sakaguchi), and samples were incubated for 90 min at 30 °C. Translation products were analysed by SDS–PAGE and gels were quantified on a Fuji FLA-3000 PhosphorImager with the use of Image Reader 8.1j software. The degree of membrane integration of each H-segment was quantified from SDS–PAGE gels by calculating an apparent equilibrium constant between the membrane-integrated and non-integrated forms: $K_{app} = f_{1g}/f_{2g}$, where $f_{1g}$ is the fraction of singly glycosylated Lep molecules and $f_{2g}$ is the fraction of doubly glycosylated Lep molecules after correcting for the fact that a fully translocated P2 domain is only about 85% glycosylated[23]. The results were then converted to apparent free energies, $\Delta G_{app} = -RT \ln K_{app}$. All $\Delta G_{app}$ values were calculated as mean values from at least two independent experiments.

**Optimization of position-specific $\Delta G_{app}$ contributions.** Position-specific residue contributions to $\Delta G_{app}$ were calculated by using an additive model with an additional term to account for the hydrophobic moment[24] ($\mu$) of the H-segment:

$$\Delta G_{app}^{pred} = \sum_{i=1}^{l} \Delta G_{app}^{aa(i)} + c_0 \sqrt{\left(\sum_{i=1}^{l} \Delta G_{app}^{aa(i)} \sin(100^\circ i)\right)^2 + \left(\sum_{i=1}^{l} \Delta G_{app}^{aa(i)} \cos(100^\circ i)\right)^2} \quad (2)$$

All amino acid profiles except those for Trp and Tyr were appoximated by single gaussians with two parameters:

$$\Delta G_{app}^{aa(i)} = a_0^{aa} e^{-a_1^{aa} i^2} \quad (3)$$

where $i$ denotes the position in the H-segment, with $i = 0$ corresponding to the central residue. To reproduce the characteristic 'W' shape of the single-scan curves for Trp and Tyr (see Supplementary Fig. 1), double gaussians (five parameters) were used for these two profiles:

$$\Delta G_{app}^{aa(i)} = a_0^{aa} e^{-a_1^{aa} i^2} + a_2^{aa}(e^{-a_3^{aa}(i-a_4^{aa})^2} + e^{-a_3^{aa}(i+a_4^{aa})^2}) \quad (4)$$

Finally, the sum of squares of the differences between measured and predicted $\Delta G_{app}$ values was minimized:

$$\hat{\Theta} = \arg \min_{\Theta} \left( \sum_{\text{All constructs}} (\Delta G_{app}^{pred} - \Delta G_{app})^2 \right) \quad (5)$$

where the sum goes over all constructs, $\Delta G_{app}$ is the experimentally measured value, $\Delta G_{app}^{pred}$ is the predicted value according to equation (2), and $\Theta = \{a_0^{aa}, a_1^{aa} \ldots a_4^{aa}, c_0\}$ is the set of 47 parameters (46 $a$ parameters needed to describe the 20 profiles according to equations (3) and (4), plus the additional hydrophobic moment weight parameter $c_0$). An underlying assumption using the above parameterization is that the profiles are symmetric around the middle of the membrane. Judging from the experimental profiles (Supplementary Fig. 1), this assumption seems justified.

A total of 321 19-residue H-segments with $\Delta G_{app}$ values between −1.5 and +1.0 kcal mol$^{-1}$ plus 3 H-segments from the Arg scan with $\Delta G_{app}$ values slightly lower than −1.5 kcal mol$^{-1}$ were used in the optimization (see Supplementary Table 1). The interval is asymmetric because the degree of double glycosylation of fully translocated H-segments can vary slightly between different batches of microsomes, introducing an extra variability in the measurements of high $\Delta G_{app}$ values.

As the starting point for the optimization, all $\Delta G_{app}^{aa(i)}$ values were set equal to zero. In a first pre-optimization step, the position-specific $\Delta G_{app}^{aa(i)}$ values were treated as being independent; that is, without the parameterization as in equations (3) and (4). The optimization was thus performed with respect to the full (20 × 19) $\overline{\Delta G_{app}^{aa}}$ matrix. Gaussian functions were then fitted to the resulting curves, and the corresponding parameter values were used as the starting point for the final optimization of the 47 parameters in $\Theta$, now using the parameterization as given by equations (3) and (4). The pre-optimization step thus served to quickly find the approximate region in parameter space for the final solution.

To estimate the ability of the model to predict $\Delta G_{app}$ values of constructs outside the training set, we performed a leave-one-out cross-validation procedure in which the model was trained on all constructs except one and then used to predict the $\Delta G_{app}$ value of the missing construct from equation (2) (Supplementary Fig. 2).

For the optimization, we used the MATLAB v. 7.0.1 (MathWorks Inc.) implementation of a subspace trust region method[25] that iteratively searches for a local minimum to the minimization criterion equation (5) based on gradient descent. The optimized $\overline{\Delta G_{app}^{aa}}$ matrix as well as the optimized $\Theta$ and length parameter values are given in Supplementary Table 2, and the resulting profiles are shown in Fig. 2.

We also tried another, previously used function[26] to fit the experimental residue-distribution profiles. The resulting profiles are similar to the gaussian profiles, and the correlation between the predicted and experimental $\Delta G_{app}$ values is essentially the same as with the gaussian profiles ($R^2 = 0.78$ versus 0.79). Because the function used in ref. 26 requires 67 parameters to describe the residue-distribution profiles (against 47 for the gaussian functions) we kept the simpler gaussian representation.

**Influence of H-segment length on $\Delta G_{app}$.** Because the derivative of $\Delta G_{app}$ with respect to segment length, $\partial \Delta G_{app}/\partial l$, increases roughly in proportion to $l$ (Supplementary Fig. 3), we assumed that a general quadratic expression accounts for the length dependence of $\Delta G_{app}$. On the basis of measured $\Delta G_{app}$ values for all 19-residue constructs analysed above but now also including the ones of variable length with $\Delta G_{app} \in [-1,1]$ kcal mol$^{-1}$ (Fig. 3), we minimized the following criterion by using the same optimization algorithm as above:

$$[\hat{c}_1, \hat{c}_2, \hat{c}_3] = \arg \min_{c_1, c_2, c_3} \left( \sum_{\text{All constructs}} (\Delta G_{app}^{pred} - \Delta G_{app} + c_1 + c_2 l + c_3 l^2)^2 \right) \quad (6)$$

where $\Delta G_{app}^{pred}$ is the predicted value according to equation (1) (that is, without consideration of length), $\Delta G_{app}$ is the measured value and $c_1$, $c_2$, and $c_3$ are parameters describing the length dependence (their optimized values are given in Supplementary Table 2). The final model used for predicting natural segments therefore contained a total of 50 parameters (47 from equation (5) plus three length parameters).

To obtain $\Delta G_{app}^{pred}$ for segments with $l \neq 19$, 'stretched' or 'compressed' $\Delta G_{app}^{aa(i)}$ profiles were used. Because the original $\Delta G_{app}^{aa(i)}$ profile values were calculated for $l = 19$, the position coordinate $j = 1, \ldots, k$ of a segment of length $k \neq 19$ was first transformed into the native coordinate system $i = -9, \ldots, +9$ (used in, for example, Fig. 2) using the expression $i = 9\{2[(j-1)/(k-1)] - 1\}$, and then the original profiles for $l = 19$ were applied.

The predicted values in Fig. 3 and Supplementary Table 1 were obtained with the leave-one-out cross-validation procedure.

**Statistical distributions of amino acids in natural transmembrane helices.** To compare the optimized $\Delta G_{app}^{aa(i)}$ profiles with statistical distributions from natural TM helices, we calculated residue distributions along the membrane normal for 575 TM helices in 158 non-redundant chains from 77 high-resolution X-ray structures[27]. Homology reduction at an 80% sequence identity threshold was performed with the CD-HIT algorithm[28]. Residue frequencies along the membrane normal were calculated and divided into bins 1.5 Å wide with respect to the distance from the membrane centre. Statistical $\Delta G_{stat}^{aa(i)}$ profiles for residue distributions in the [−45, +45] Å distance interval were then calculated as

$$\Delta G_{stat}^{aa(i)} = -RT \ln \left( \frac{f(aa,i)}{bgr(aa)} \right) \quad (7)$$

where $f$(aa,$i$) is the frequency of amino acid aa in helix position $i$ normalized such that the sum over all amino acids adds to 1, and bgr(aa) is the background frequency of amino acid aa, according to the amino acid composition of SwissProt[29] (version 50.1), resembling the procedures used in similar studies[26,30]. Finally, to smooth the profiles, least-squares curve fitting was performed in accordance with equations (3) and (4). To compare the profiles with the $\overline{\Delta G_{app}^{aa}}$ matrix it was assumed that each amino acid in the H-segments corresponded to a $z$-coordinate displacement of 1.5 Å. Although the curve fitting was performed with respect to the full [−45, +45] Å interval, in Fig. 2 only the [−13.5, +13.5] Å interval of the curves is shown.

**Distributions of $\Delta G_{app}^{pred}$ values in natural proteins.** To investigate how the distributions of $\Delta G_{app}^{pred}$ values differ between sets of secreted, transmembrane and cytoplasmic proteins, we compiled four data sets that were all homology-reduced at an 80% sequence identity threshold with the CD-HIT algorithm[28]: first, all mammalian proteins in SwissProt[29] (version 50.1) annotated to have a signal peptide and exactly one transmembrane region (349 single-spanning membrane proteins); second, all mammalian proteins in SwissProt annotated to have a signal peptide but no transmembrane regions (1,012 soluble proteins targeted to the secretory pathway); third, all mammalian proteins in SwissProt annotated with 'cytoplasm' as subcellular location (670 cytoplasmic proteins); and fourth, all known X-ray structures of membrane proteins from the OPM database[27] with at least two TM helices (66 multi-spanning membrane proteins with a total of 508 TM helices). Proteins annotated in SwissProt as having a GPI anchor were removed before the analysis, because they were in

some cases annotated as having a transmembrane segment and in some cases not. Annotated signal peptides were removed from all sequences.

A sliding-window approach was employed to identify the segment of length 17–33 residues with the lowest $\Delta G_{app}^{pred}$. In the first to third sets we scanned the full protein sequences, whereas in the fourth set we extended each annotated TM helix by ten residues on both the amino-terminal end and the carboxy-terminal end and then scanned for such a segment.

To compare the $\Delta G_{app}^{pred}$ predictions against existing hydrophobicity-based predictions, a similar sliding-window analysis, but with a fixed window length ($l = 19$), was also performed with the Zhao–London[31], Kyte–Doolittle[32] and Wimley–White[33] hydrophobicity scales. Because these scales do not contain position-dependent information, for comparison we also made predictions with a simpler version of the 'biological' scale, in which all profiles were replaced with their respective mean $\Delta G_{app}^{aa(i)}$ value (that is, no positional dependence), and in addition the terms modelling length and hydrophobic moment were left out ($c_0$, $c_1$, $c_2$ and $c_3$ were set to zero, and the window length was fixed at $l = 19$). The resulting distributions are shown in Supplementary Fig. 6. If $l$ is allowed to vary between 17 and 33 residues (as in the $\Delta G_{app}^{pred}$ calculations), the overlaps between the distributions for single-spanning and multi-spanning membrane change only slightly (data not shown).


23. Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433,** 377–381 (2005).
24. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179,** 125–142 (1984).
25. Coleman, T. F. & Li, Y. An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optimiz.* **6,** 418–445 (1996).
26. Senes, A. *et al.* $E_z$, a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: Derivation and applications to determining the orientation of transmembrane and interfacial helices. *J. Mol. Biol.* **366,** 436–448 (2007).
27. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: orientations of proteins in membranes database. *Bioinformatics* **22,** 623–625 (2006).
28. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
29. O'Donovan, C. *et al.* High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **3,** 275–284 (2002).
30. Ulmschneider, M. B., Sansom, M. S. & Di Nola, A. Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins* **59,** 252–265 (2005).
31. Zhao, G. & London, E. An amino acid 'transmembrane tendency' scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity. *Prot. Sci.* **15,** 1987–2001 (2006).
32. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132 (1982).
33. Wimley, W. C., Creamer, T. P. & White, S. H. Solvation energies of amino acid sidechains and backbone in a family of host–guest pentapeptides. *Biochemistry* **35,** 5109–5124 (1996).