How many membrane proteins are there?

DANA BOYD, CLARK SCHIERLE, AND JON BECKWITH
Department of Molecular Genetics and Microbiology, Harvard Medical School, Boston, Massachusetts 02115
(RECEIVED September 10, 1997; ACCEPTED October 22, 1997)

Abstract
One of the basic issues that arises in functional genomics is the ability to predict the subcellular location of proteins that are deduced from gene and genome sequencing. In particular, one would like to be able to readily specify those proteins that are soluble and those that are inserted in a membrane. Traditional methods of distinguishing between these two locations have relied on extensive, time-consuming biochemical studies. The alternative approach has been to make inferences based on a visual search of the amino acid sequences of presumed gene products for stretches of hydrophobic amino acids. This numerical, sequence-based approach is usually seen as a first approximation pending more reliable biochemical data. The recent availability of large and complete sequence data sets for several organisms allows us to determine just how accurate such a numerical approach could be, and to attempt to minimize and quantify the error involved. We have optimized a statistical approach to protein location determination. Using our approach, we have determined that surprisingly few proteins are misallocated using the numerical method. We also examine the biological implications of the success of this technique.

Keywords: computer modeling; discriminant analysis; hydropathy; membrane proteins; statistical methods

Experimental methods of determination of subcellular protein location are accurate but time-consuming. Hydropathy analysis (Kyte & Doolittle, 1982) has often been used to deduce subcellular localization of proteins in the absence of experimental data. However, although visual inspection of hydropathy plots can be useful in predicting the topology of known integral membrane proteins, it is ineffective as an accurate predictor of the location of a protein. To discriminate between integral and peripheral membrane proteins, Klein et al. (1985) generated a single number, maxH, the average hydropathy of the most hydrophobic protein segment of a given length for a given protein using a given hydropathy scale. In the interest of clarity, we will refer to this number as the "maxH value," while using the term "maxH segment" to refer to the hydrophobic peptide segment to which it belongs. This was then applied to a set of known integral and peripheral membrane proteins in a training set. A discriminator function was generated that assigned a probability of being an integral membrane protein to a given value of maxH. This function was then used to analyze a similar set of known proteins, the test set. It was determined that the Kyte-Doolittle hydropathy scale and a window length of 17 residues gave the best resolution of membrane and soluble proteins in the test set. We have found that the method used by Klein et al. (1985) is still generally useful, but that the actual functions provided in their paper are not, having been derived at a time when very few proteins were both sequenced and characterized.

Results
We discovered the need for a new discriminator when we attempted to apply the functions described by Klein et al. (1985) to the Escherichia coli genome. We used as a cutoff the maxH value proposed by Klein et al. at which it was equally probable that a protein is in the membrane or not and found that many soluble E. coli proteins were misclassified as membrane proteins. The converse was not the case. This result suggested that a new discriminator function, trained on E. coli proteins, was needed. We used the November 1996 SWISSPROT release 34 E. coli file annotations to identify E. coli proteins with known subcellular location. We excluded all proteins with terms such as "hypothetical," "putative," or "possible" in their identification fields and those with names starting with "Y," indicating an open reading frame. This gave us a group of 397 proteins with subcellular locations listed as cytoplasmic (171), periplasmic (66), or integral inner membrane (160). (Proteins with other subcellular locations including outer membrane proteins or membrane-associated proteins were not included at this stage.) These were segregated randomly into training and test sets. The maxH value was calculated for each protein in the training set using a variety of different hydropathy scales and eight window lengths (odd values from 7 to 21). Results with five of the hydropathy scales are presented in the tables that follow: KD, Kyte and Doolittle (1982); EI, Eisenberg (1984); GvH, von Heijne (1992); GES, Engelmann et al. (1985); and a new scale, JTT2, which is given in Table 1. Normal distribution parameters (mean and SD, for each subgroup in the test set, cytoplasmic, periplasmic, and integral as well as cytoplasmic + periplasmic) were calculated. For many scales and window lengths, the means
Table 1. The JTT2 scale

<table>
<thead>
<tr>
<th>A</th>
<th>1.37</th>
<th>G</th>
<th>1.03</th>
<th>M</th>
<th>1.39</th>
<th>S</th>
<th>0.83</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1.12</td>
<td>H</td>
<td>0.74</td>
<td>N</td>
<td>0.43</td>
<td>T</td>
<td>0.89</td>
</tr>
<tr>
<td>C</td>
<td>0.17</td>
<td>I</td>
<td>2.2</td>
<td>P</td>
<td>0.51</td>
<td>V</td>
<td>1.81</td>
</tr>
<tr>
<td>D</td>
<td>0.16</td>
<td>K</td>
<td>0.19</td>
<td>Q</td>
<td>0.35</td>
<td>W</td>
<td>1.56</td>
</tr>
<tr>
<td>E</td>
<td>1.93</td>
<td>L</td>
<td>1.78</td>
<td>R</td>
<td>0.3</td>
<td>Y</td>
<td>1.01</td>
</tr>
</tbody>
</table>

*JTT2 is a relative frequency scale. The values are a ratio, the frequency with which each amino acid occurs in transmembrane segments (Jones et al., 1992) divided by the overall frequency of each amino acid in a large set of proteins (Jones et al., 1994).*

of cytoplasmic and periplasmic proteins were close together, whereas the mean for integral membrane proteins was separated by several standard deviations from the others. The periplasmic and cytoplasmic proteins were therefore considered as one group, soluble proteins.

The maxH value at which it is equally likely that a protein is integral or not was found for each scale and window length by finding the point at which the normal density functions for each group, integral or not, derived from the training set data were equal (Table 2). The tester set was then screened with each of these discriminator values to determine how well each scale and window length performed (Table 3). All scales performed more poorly at short window lengths. At lengths from 15 to 21, the E1 (with four or five misallocations) and JTT2 (three to six misallocations) scales performed best. The KD scale was nearly as good as those mentioned above. The performance of the scales in this test generally correlates with the number of standard deviations (taken as the average of standard deviations for both classes, soluble and integral membrane) separating the means (data not shown). The proteins with subcellular locations given in SwissProt other than the three classes used in the analysis included those classified as peripheral membrane, membrane-associated, membrane-bound, outer membrane lipoprotein, and a few other types. Among these, seven have high maxH values and are clearly polytopic integral membrane proteins. The remaining 93, including all outer membrane proteins, have maxH values below the discriminator value.

The scale JTT2 and window length 19 was taken for subsequent analysis, but the other scales mentioned above could also be used at long window lengths. (This scale was chosen because it gives the best separation between peaks in the training set and is equivalent at this length to the E1 scale, Table 3.) The maxH value was calculated for all proteins in the E. coli genome. The results are shown as a histogram with the maxH values on the horizontal axis and the number of proteins with each value on the vertical axis at the top left of Figure 1. The distribution of proteins is clearly bimodal. The proteins are separated into two groups, one with a mean maxH value of 1.32 and another with mean maxH of 1.61. The nadir between the two peaks corresponds closely to the discriminator value determined from the training set, 1.505 (shown as a hatch mark on the horizontal axis in Fig. 1), which was determined from the subset whose location is known. We propose that the group of proteins with maxH values above the discriminator value is predominantly integral membrane proteins and that the fraction of E. coli proteins in this group, 21%, is a good estimate of the number of membrane proteins in E. coli.

Table 2. Discriminator values and quadratic discriminator function

<table>
<thead>
<tr>
<th>Window length</th>
<th>KD</th>
<th>EI</th>
<th>GvH</th>
<th>GES</th>
<th>JTT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>1.98</td>
<td>0.286</td>
<td>0.301</td>
<td>1.84</td>
<td>1.48</td>
</tr>
<tr>
<td>19</td>
<td>2.1</td>
<td>0.308</td>
<td>0.357</td>
<td>2.02</td>
<td>1.51</td>
</tr>
<tr>
<td>17</td>
<td>2.25</td>
<td>0.332</td>
<td>0.411</td>
<td>2.15</td>
<td>1.53</td>
</tr>
<tr>
<td>15</td>
<td>2.44</td>
<td>0.36</td>
<td>0.462</td>
<td>2.26</td>
<td>1.57</td>
</tr>
<tr>
<td>13</td>
<td>2.56</td>
<td>0.381</td>
<td>0.51</td>
<td>2.35</td>
<td>1.6</td>
</tr>
<tr>
<td>11</td>
<td>2.79</td>
<td>0.407</td>
<td>0.563</td>
<td>2.48</td>
<td>1.64</td>
</tr>
<tr>
<td>9</td>
<td>3.04</td>
<td>0.439</td>
<td>0.62</td>
<td>2.62</td>
<td>1.69</td>
</tr>
<tr>
<td>7</td>
<td>3.38</td>
<td>0.483</td>
<td>0.722</td>
<td>2.81</td>
<td>1.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sc</th>
<th>W</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>50%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>KD</td>
<td>19</td>
<td>4.272815</td>
<td>-40.73497</td>
<td>66.727646</td>
<td>2.101</td>
<td>2.236</td>
</tr>
<tr>
<td>KD</td>
<td>17</td>
<td>3.239692</td>
<td>-34.560265</td>
<td>61.458291</td>
<td>2.255</td>
<td>2.409</td>
</tr>
<tr>
<td>EI</td>
<td>21</td>
<td>232.34478</td>
<td>-286.771384</td>
<td>63.048289</td>
<td>0.286</td>
<td>0.306</td>
</tr>
<tr>
<td>EI</td>
<td>19</td>
<td>304.795979</td>
<td>-350.066027</td>
<td>79.151489</td>
<td>0.308</td>
<td>0.327</td>
</tr>
<tr>
<td>JTT2</td>
<td>19</td>
<td>94.724066</td>
<td>-420.066027</td>
<td>417.682121</td>
<td>1.505</td>
<td>1.528</td>
</tr>
<tr>
<td>JTT2</td>
<td>15</td>
<td>123.133257</td>
<td>-501.711424</td>
<td>484.481094</td>
<td>1.572</td>
<td>1.600</td>
</tr>
</tbody>
</table>

*The maxH value at which it is equally likely that a protein with that maxH value is integral or not in the training set is given for each scale and window length.*

*Values of a, b, and c in the following expression: \( e^{ax_{\text{maxH}}^2 + bx_{\text{maxH}} + c} \), which gives the odds that a protein is not integral. Also given are the values of maxH at which it is 50% and 95% probable that a given protein is integral. Sc, scale; W, window length.*
Table 3. Test set misallocations

<table>
<thead>
<tr>
<th>Window length</th>
<th>KD</th>
<th>EI</th>
<th>Scale GvH</th>
<th>GES</th>
<th>JTT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>8</td>
<td>4</td>
<td>9</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>19</td>
<td>11</td>
<td>4</td>
<td>11</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>5</td>
<td>11</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>5</td>
<td>12</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>9</td>
<td>12</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>15</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>11</td>
<td>19</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>10</td>
<td>32</td>
<td>12</td>
<td>11</td>
</tr>
</tbody>
</table>

*A quadratic discriminator similar to that of Klein et al. (1985) was obtained by setting the normal density functions for the training set periplasmic + cytoplasmic proteins and integral proteins equal to each other after multiplication by a normalization factor to account for the different numbers of proteins in the two groups. The number of misallocated test set proteins for each discriminator value is given. The number of averaged standard deviation separating the two classes of proteins in the training set is in each case close to the minimum point between the two peaks.

We have used the same approach to classify all the proteins in four other genomes for which there is complete data, *Haemophilus influenza*, *Mycoplasma genitalium*, *M. pneumoniae*, and *Methanococcus jannaschii*. Four additional genomes, for which partial data is available in relatively nonredundant sets, were also examined, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens*. The maxH value was calculated and plotted as for *E. coli*. The discriminating maxH determined from *E. coli* is represented by a vertical tick below the horizontal axis (Fig. 1). The results are quite similar to those seen for *E. coli*. In each case, there is a bimodal distribution. The discriminator value is in each case close to the minimum point between the two peaks. If the assumption is made that the peak with a higher average hydrophathy consists primarily of membrane proteins as in *E. coli*, the predicted number of integral membrane proteins varies between 12 and 20% of the total in the different organisms for which whole-genome data are available (Table 4). We propose that this is a reasonable estimate of the number of membrane proteins in each of these organisms. Note that these values are significantly different from values reported by other authors in their analyses of genomes (Tomb et al., 1997).

The success of the maxH method to discriminate between soluble and integral membrane proteins is intriguing. Recall that the maxH value represents the average hydrophathy of the single most hydrophobic segment (the maxH segment) of the protein in question. Note also that the best discriminators use window lengths that approximate the length of a membrane-spanning segment. We have considered two explanations for the significance of the maxH value. First, it may merely reflect the lowest average hydrophathy for a protein segment of appropriate length that would permit that protein to be stably integrated into a membrane. This explanation would suggest that, as we examine proteins with increasing numbers of transmembrane segments, the average value of maxH would increase. That is, the probability is that a multi-spanning membrane protein will have a sequence with a higher maxH than that of a protein that spans the membrane only once. Alternatively, the maxH value may correspond to a transmembrane segment with particularly high hydrophobicity, which is essential for any protein to attain a membrane location. If this were the case, the maxH value would remain more or less the same whether we examined bitopic membrane proteins or multi-spanning membrane proteins.

We began investigating this issue by classifying the set of known membrane proteins into groups based on the number of transmembrane segments that they possess. We then calculated the average of the maxH values in each group, as well as the average hydropaths of all the other transmembrane segments in the proteins in the group. If the maxH segment is indistinguishable from the other transmembrane segments, the average values of maxH should increase with increasing numbers of transmembrane segments (Fig. 2A). If the value of maxH is unique, however, then the values should remain relatively constant, reflecting a higher average hydrophathy than that required of the other transmembrane segments.

We tested these hypotheses by generating artificial genomes using different rules for constructing the transmembrane proteins and comparing data from these artificial genomes with the real *E. coli* data set (Fig. 2B). The first rule randomized the transmembrane domains, mixing the maxH segments with the rest. This yielded a strong positive correlation of maxH value with increasing number of transmembrane segments, and was statistically highly significant.
Fig. 2. Correlation of maxH and number of transmembrane helices. Two models for the significance of maxH are presented. A: The distributions of maxH values are represented by the shaded area, and the larger distributions represent all transmembrane segments. If the maxH value were not unique, then proteins with more transmembrane helices would have higher maxH values as indicated in the case labeled “Random.” If, on the other hand, there were some requirement for integral proteins to have at least one transmembrane helix with higher than average hydropathy, then there would be less difference between proteins with few and many transmembrane helices, as reflected by the line labeled “Unique.” B: Actual data from transmembrane proteins of E. coli and averaged data from several simulated genomes are shown. The solid line indicates the actual trend of maxH values for E. coli transmembrane proteins with increasing number of transmembrane segments. The other lines represent averaged data from artificial genomes generated by one of several rules: Rule 0: Average hydropathy values for all transmembrane segments in E. coli were placed in a pool and a new set of artificial transmembrane proteins were generated by selecting values at random from the pool. Rule 1: Artificial genomes were generated as with rule 0, except all maxH values were placed in a separate pool, and each artificial membrane protein was generated by first selecting one value from the maxH pool. Rule 2: Artificial genomes were generated from a single pool as with rule 0, but the maxH segment of each protein was required to have a hydropathy greater than the discriminator value for E. coli. Rule 3: Artificial genomes were again generated using a single starting pool, but the maxH segment for each artificial membrane protein was generated using the discriminator function for E. coli.

Table 4. MaxH genomic means and standard deviations

<table>
<thead>
<tr>
<th>Species</th>
<th>First peak</th>
<th></th>
<th>Second peak</th>
<th></th>
<th>Second peak as % of total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>E. coli</td>
<td>1.31</td>
<td>0.11</td>
<td>1.61</td>
<td>0.07</td>
<td>21%</td>
</tr>
<tr>
<td>H. inf.</td>
<td>1.31</td>
<td>0.11</td>
<td>1.61</td>
<td>0.07</td>
<td>19%</td>
</tr>
<tr>
<td>M. jann.</td>
<td>1.31</td>
<td>0.11</td>
<td>1.64</td>
<td>0.09</td>
<td>17%</td>
</tr>
<tr>
<td>M. pneu.</td>
<td>1.33</td>
<td>0.14</td>
<td>1.68</td>
<td>0.07</td>
<td>17%</td>
</tr>
<tr>
<td>M. gen.</td>
<td>1.34</td>
<td>0.14</td>
<td>1.69</td>
<td>0.07</td>
<td>20%</td>
</tr>
<tr>
<td>Human</td>
<td>1.31</td>
<td>0.21</td>
<td>1.58</td>
<td>0.09</td>
<td>ND</td>
</tr>
<tr>
<td>C. ele.</td>
<td>1.31</td>
<td>0.15</td>
<td>1.57</td>
<td>0.10</td>
<td>ND</td>
</tr>
<tr>
<td>Yeast</td>
<td>1.32</td>
<td>0.17</td>
<td>1.54</td>
<td>0.11</td>
<td>ND</td>
</tr>
</tbody>
</table>

*The mode was determined and a mean and standard deviation were calculated from the outer halves of the data. The percent of the total number of proteins in the second peak was calculated as the fraction with maxH values above 1.505. Correcting the data for overlap between the two peaks gave an estimate of 849 integral membrane proteins in E. coli, 20% of the total 4,280 used in the calculation.

*ND, not done.

*Mode not determined accurately for this peak.
How many membrane proteins are there?

unlikely to reflect the real E. coli data (p < 0.002%). A rule which separated maxH segments into a separate pool prior to randomization, thus treating them as unique, yielded data statistically indistinguishable from E. coli (p = 20%). Two other rules that generated new maxH segments using either a discriminator cutoff or a discriminator function yielded data even more similar to E. coli (p = 60%). Thus, it is highly improbable that the pattern of maxH values seen in E. coli reflects a random assortment of hydrophobic residues among all transmembrane segments. Rather, the situation in E. coli more closely resembles simulations in which maxH segments are given a unique, higher hydropathy distribution than the other transmembrane segments.

Discussion

We have presented an optimized scheme for numerical discrimination between membrane localized and soluble proteins. Using rigorous statistical techniques and improved data sets, very high levels of reliability have been achieved for the prediction of the subcellular location of a given protein. Using this technique, it should be possible to develop models of strong predictive value for the growing number of organisms for which large data sets are available. Our own preliminary analysis of different organisms as presented here indicates that there is some variation in the exact values for the means in different organisms. The histograms for both Mycoplasma species reach minima between the two peaks at a hydropathy that is above 1.51 (Fig. 1). In these two species, the mean of the maxH values in the second peak is higher than in the other bacteria (Table 3). A higher maxH discriminator would be appropriate for these organisms. (A value of 1.56 can be obtained from the data with the assumption that the two peaks actually represent integral and nonintegral proteins.) The distribution of maxH values in the three eukaryotes examined, although clearly similar to that in bacteria and archa, is not always so cleanly separated into two populations. These differences may represent differences in the properties of the membranes or membrane localization mechanisms of the eukaryotes, bacteria, and archa. For instance, lipid composition of membrane compartments from members of the three kingdoms may differ, resulting in different requirements for optimal membrane insertion properties.

In addition, we have addressed the possible biological significance of the success of the maxH method of discrimination. Our modeling demonstrates that the maxH segment is statistically distinguishable from other transmembrane segments. These results are not meant to imply that we have determined the rule by which E. coli membrane proteins are recognized and localized. Whether this reflects a mechanism by which membrane proteins are targeted to the membrane or some evolutionary optimization of membrane insertion and stability is unclear. We can, however, conclude that the value of maxH is uniquely hydrophobic among transmembrane segments, and that this may explain why the discriminator is as successful as it is.

Materials and methods

Sequence data were obtained from the public domain (GenBank, PIR, SwissProt, TMBase, PDB). Numerical discriminant analysis was performed as described in Anderson (1958) and Kendall and Stuart (1976).

Acknowledgments

We thank Don Engelman, Eric Stewart, and Mark Gerstein for advice and encouragement. This work was supported by a grant from the National Institutes of Health (GM54160).

References