Prediction of Protein Inter-domain Linker Regions by a Hidden Markov Model

Kyounghwa Bae¹, Bani K. Mallick¹ and Christine G. Elsik²

¹Department of Statistics, Texas A&M University, College Station, TX, 77843-3143, USA and ²Department of Animal Science and Intercollegiate Faculty of Genetics, Texas A&M University, College Station, TX, 77843-2471, USA

ABSTRACT

Motivation: We wish to predict protein inter-domain linker regions using sequence alone, without requiring known homology. Identifying linker regions will delineate domain boundaries, and can be used to computationally dissect proteins into domains prior to clustering them into families. We develop a hidden Markov model (HMM) of linker/non-linker sequence regions using a linker index derived from amino acid propensity. We employ an efficient Bayesian estimation of the model using Markov Chain Monte Carlo (MCMC), particularly Gibbs sampling, to simulate parameters from the posteriors. Our model recognizes sequence data to be continuous rather than categorical, and generates a probabilistic output.

Results: We applied our method to a dataset of protein sequences in which domains and inter-domain linkers had been delineated using the Pfam-A database. The prediction results are superior to a simpler method that also uses linker index.

Contact: c-elsik@tamu.edu; bmallick@stat.tamu.edu; kbae@stat.tamu.edu

Supplementary Information: http://racerx00.tamu.edu/kbae

INTRODUCTION

The fundamental unit of protein structure, the domain, is defined as a unit that can independently fold into a stable tertiary structure. Domains often evolve as independent units that are found in different combinations. Thus, the domain has alternatively been defined as an evolutionary unit. Domain identification within a protein sequence is valuable in numerous applications. It allows structural determination of separate domains, which is often more successful than solving whole proteins. Computational methods for clustering proteins based on sequence similarity perform better when sequences are fragmented into single domain units.

Domain boundary prediction methods that apply the structural definition consider the domain to be a compact, semi-independent unit with a hydrophobic core; these methods use atomic coordinates from experimentally determined 3-dimensional structures (e.g. Holm and Sander, 1994; Islam et al., 1995; Siddiqui and Barton, 1995; Wernisch et al. 1999; Taylor 1999) or predicted structure (George and Heringa, 2002c; Marsden et al., 2002). Other methods apply the the evolutionary definition of domain, and use regions of conservation in sequence alignments to identify domain boundaries (e.g. Sonnhammer and Kahn 1994; Gouzy et al., 1997; Gracy and Argos, 1998; George and Heringa, 2002b). The domain families in Pfam-A are created using profile hidden Markov models built on multiple sequence alignments (Bateman et al., 2004). The CHOP method cuts proteins into domain-like fragments using domain boundary information from proteins with known structure and from Pfam-A (homology-based) domains (Liu and Rost, 2004a). CHOPNet is a neural network method that does not rely on known homology or known structure, but uses as input both evolutionary information and predicted structure (Lui and Rost, 2004b). The DGS (Domain Guess by Size) method makes domain boundary estimates based on the statistical distribution of protein and domain lengths in a representative set (Wheelan et al., 2000).

An alternative to delineating domain boundaries is identifying inter-domain linkers. The linker is defined as a region between adjacent domains. Studies have shown that linkers can play an essential role in maintaining cooperative inter-domain interactions (Gokhale and Khosla, 2000). An understanding of linker properties will aid the engineering of fusion proteins. The composition and length of linkers have been shown to affect protein stability, folding and domain-domain orientation (e.g. Robinson and Sauer, 1998). As the alternative to domain prediction, linker identification facilitates splitting multidomain proteins into single domains prior to structural analysis or computational protein clustering.

Studies of linkers in various protein families have shown that linker regions lack regular secondary structure (e.g. Argos, 1990), but a recent study has identified helical linkers (George and Heringa 2002a). Studies agree that some amino acids are more prevalent in the linker regions than in the domain regions (Robinson...
and Sauer, 1998; George and Heringa, 2002a; Tanaka et al., 2003). Most methods for identifying linkers use predicted secondary structure, amino acid propensity, or a combination of the two. Miyazaki et al. (2002) have applied a neural network to predict the linker boundaries based on amino acid propensity, and found that linkers possess characteristics that may distinguish them from intra-domain loops. The method of Tanaka et al. (2003) combines predicted secondary structure with amino acid propensity to identify loop regions and distinguish linker and non-linker loops. The Udwary-Merski algorithm (Udwary et al., 2002) combines three properties of linkers: low sequence conservation identified by multiple sequence alignment, low secondary structure conservation and low hydrophobicity. DomCut, which predicts linker regions based on sequence alone, relies solely on amino acid propensity (Suyama and Ohara, 2003). This method simply defines a linker region to be one that has lower linker index values than a specified threshold value. Similar to the approach taken in DomCut, we will use linker index to model linker regions, and will apply our model to a dataset of evolutionarily defined domains. We will employ a hidden Markov model to predict not only linker regions, but also their boundaries.

Hidden Markov models (HMMs) have been employed in diverse areas of computational biology (e.g. Lander and Green, 1987; Churchill, 1989; Cardon and Stormo, 1992; Burge and Karlin, 1997). Krogh et al. (1994) applied HMMs to multiple alignment of protein families and domains. Asai et al. (1993) applied HMMs to protein secondary structure prediction. Later, Schmidler et al. (2000, 2001) applied generalized HMMs with Bayesian estimation to protein secondary structure prediction. The observations in the HMMs for protein structure prediction are recognized as strings of amino acids (categorical variables), forming the primary sequence of a protein.

In this paper, sequences are assumed to have a structure composed of regions that are homogeneous within a region but may differ between regions. We assume that protein sequence data is produced by a hidden Markov model and compositional variation is likely to reflect functional or structural differences between regions. Each region is classified into one of a finite number of states (linker and non-linker); we wish to estimate the states given the observed protein sequence. Importantly, we recognize the protein sequence data as continuous data instead of categorical data, which is an alternative approach to most HMMs in computational protein sequence analysis. Instead of recognizing the protein sequence as a string of amino acids (categorical variables), we recognize the protein sequence as a string of linker index values (continuous variables). Our objective is to identify linker index values that discriminate linker and non-linker regions.

Parameter estimation in HMMs usually relies on maximum likelihood or the Bayesian approach. In the Bayesian approach, we consider the HMM as a mixture model with missing data. We can associate observation \( y_i \) with missing data \( z_i \), which represents the state (e.g. linker or non-linker) from which \( y_i \) is generated (Robert and Mengersen, 1999). An important feature of our method is that we overcome the problem of missing data by employing an efficient Bayesian estimation of the model through a Markov Chain Monte Carlo (MCMC) method, particularly Gibbs sampling (Gelfand and Smith, 1992; Gilks et al., 1996). Other methods of handling missing data have been the use of the EM algorithm (Dempster et al., 1977) or a recurrent forward-backward formula. The EM algorithm was originally tailored for missing data structures, but dependency between states causes problems in mixture estimation. While the simulation of the missing data is straightforward for an independent structure, it is quite difficult to simulate from the distribution of missing data that is conditional on the observed data in HMMs. The use of a recurrent forward-backward formula, which is widespread in the literature for estimating HMM parameters, is time consuming and numerically sensitive. Instead, our method uses Gibbs sampling, which effectively reduces the problem of sampling from a high-dimensional distribution to sampling from a series of low-dimensional distributions.

DATA

Data preparation

We downloaded protein sequence data from the Pfam database release 14 (Bateman et al., 2004) to construct a representative dataset of multidomain protein sequences. Pfam-A is a collection of domain families created using profile HMMs built on multiple alignments of homologous proteins. Release 14 of the Pfam database contains protein sequences from Swiss-Prot release 43.2 and SP-TrEMBL release 26.2 (Boeckmann et al. 2003). The Pfam database provides protein sequence coordinates for Pfam-A domains identified in these proteins. Protein sequences that were annotated as containing transmembrane regions in the Pfam database were removed from the dataset. We define a linker as a sequence segment of 4 to 20 residues that connects two adjacent regions identified by Pfam as domains. The reasoning behind this length range is that an inter-domain segment longer than 20 residues may contain a domain that has not yet been identified, instead of being one long linker region. We also define non-linker regions as sequence segments excluding linker regions. We denote a whole sequence as Full. We used only protein sequences whose entire length can be classified as linker or domain by our criteria, except we allowed up to 20 non-domain residues at the N- and C-termini. By this procedure, we obtained 11968...
sequences with at least one linker region (14339 linker, 28726 corresponding domains and 824 unique domain regions).

We removed redundancy in this dataset as follows. First, we grouped the 11968 proteins into homeomorphic families (identical domain organization). We performed an all by all sequence comparison of the 11968 sequences using FASTA (Pearson and Lipman, 1988). We then applied single-linkage clustering using criteria of E-value ≤ 10^{-6} and at least 80% alignment coverage. Some of the resulting clusters contained sequences with different domain organizations, due to the transitive nature of single-linkage clustering. Therefore, instead of selecting only one sequence from each cluster, we selected one sequence from each domain organization within each cluster. We also removed 7 protein sequences which were significantly longer than the rest (> 1000 residues).

We obtained 802 sequences with at least one linker region. These 802 sequences contained 993 linkers and 1988 corresponding domain regions from 376 unique Pfam-A domain families. The average length of linkers and domains was 11.24 and 141.38, respectively. The relative frequency of individual amino acids were compared between linker region and other regions by a z-test. Amino acid whose frequency is significantly different between linker and domain (P-value < 10^{-3}) are indicated by (*) in Table 1. The distribution of amino acids in the linker database of George and Heringa (2002a) shows similar patterns even though the definition of linker region is different. We can incorporate the difference in amino acid composition among regions into our model using the linker index.

**Linker index**

Many studies have reported observations of some amino acids at higher frequency in the linker regions than in the domain regions. Proline (P), Lysine (K), Glutamic acid (E), Serine (S), Aspartic acid (D) and Glutamine (Q) are preferred amino acids in linker regions. Studies (George and Heringa, 2002a; Suyama and Ohara, 2003; Tanaka et al., 2003) have shown Proline to be the most preferred linker amino acid. However, there is disagreement among studies regarding the other preferred linker amino acids. It is no surprise that Proline is favored because it has no amide hydrogen to donate in hydrogen bonding, and therefore structurally isolates the linker from domains (George and Heringa, 2002a). The analysis of our dataset also shows that Proline is the most preferred amino acid in the linker regions.

The propensity of amino acids for linkers have been determined by other groups in three ways: 1) by comparing linker regions to domain regions (Suyama and Ohara, 2003), 2) by comparing linker regions to all non-linker regions (domains and terminal sequence; Tanaka et al., 2003) and 3) by comparing linker regions to Full sequences (linkers, domains and terminal sequences; George and Heringa, 2002a). We found amino acid frequencies to be similar among domains, non-linkers and Full sequences (see supplementary website), so we proceeded using amino acid propensity for linkers compared to domains.

To incorporate the difference in amino acid composition between domain and linker regions, we employ the linker index, \( y_l \), which reflects the preference of amino acids in the linkers relative to the domain region, from Suyama and Ohara (2003).

\[
y_l = -\ln \left( \frac{f_{l\text{ linker}}}{f_{l\text{ domain}}} \right)
\]

where \( f_{l\text{ linker}} \) is the relative frequency of the amino acid \( l \) in the linker (domain) region in the data set. Because \( y_l \) represents the preference for amino acid \( l \) in the linker region, we note that the value of \( y_l \) will be negative if the relative frequency of amino acid \( l \) in the linker region is greater than its relative frequency in the domain region.

To calculate the smoothed linker index, we took an average of the linker index within each window size \( \omega \) and assigned this averaged linker index value \( y \) to the center amino acid of the window by sliding from

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Linker Frequency</th>
<th>Domain Frequency</th>
<th>( y_l )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7.97 (7.94)</td>
<td>8.10</td>
<td>0.0166</td>
</tr>
<tr>
<td>C</td>
<td>0.89 (1.24)</td>
<td>1.50*</td>
<td>0.5724</td>
</tr>
<tr>
<td>D</td>
<td>6.32 (5.28)</td>
<td>5.60*</td>
<td>-0.1278</td>
</tr>
<tr>
<td>E</td>
<td>7.97 (6.89)</td>
<td>6.60*</td>
<td>-0.1794</td>
</tr>
<tr>
<td>F</td>
<td>2.74 (4.34)</td>
<td>4.03*</td>
<td>0.3561</td>
</tr>
<tr>
<td>G</td>
<td>7.74 (6.14)</td>
<td>7.37</td>
<td>-0.0442</td>
</tr>
<tr>
<td>H</td>
<td>1.91 (2.32)</td>
<td>2.27</td>
<td>0.1643</td>
</tr>
<tr>
<td>I</td>
<td>4.73 (5.13)</td>
<td>6.37*</td>
<td>0.2758</td>
</tr>
<tr>
<td>K</td>
<td>6.97 (5.72)</td>
<td>5.81*</td>
<td>-0.2134</td>
</tr>
<tr>
<td>L</td>
<td>7.51 (9.60)</td>
<td>9.54*</td>
<td>0.2523</td>
</tr>
<tr>
<td>M</td>
<td>2.13 (2.15)</td>
<td>2.24</td>
<td>0.0197</td>
</tr>
<tr>
<td>N</td>
<td>4.22 (4.12)</td>
<td>4.08</td>
<td>-0.0786</td>
</tr>
<tr>
<td>P</td>
<td>6.63 (6.07)</td>
<td>4.30*</td>
<td>-0.4188</td>
</tr>
<tr>
<td>Q</td>
<td>3.90 (4.05)</td>
<td>3.33*</td>
<td>-0.1051</td>
</tr>
<tr>
<td>R</td>
<td>5.77 (5.79)</td>
<td>5.24</td>
<td>-0.0762</td>
</tr>
<tr>
<td>S</td>
<td>7.20 (5.55)</td>
<td>6.13*</td>
<td>-0.1629</td>
</tr>
<tr>
<td>T</td>
<td>5.80 (5.66)</td>
<td>5.35</td>
<td>-0.0701</td>
</tr>
<tr>
<td>V</td>
<td>6.24 (6.64)</td>
<td>7.34*</td>
<td>0.1782</td>
</tr>
<tr>
<td>W</td>
<td>0.81 (1.24)</td>
<td>1.32*</td>
<td>0.3836</td>
</tr>
<tr>
<td>Y</td>
<td>2.46 (3.47)</td>
<td>3.38*</td>
<td>0.2500</td>
</tr>
</tbody>
</table>

Table 1. Amino acid frequency in the different regions of the protein and the linker index of amino acids.
the N-terminus to the C-terminus of a protein sequence. We used a window size, \( w = 9 \), which provided the greatest discrimination between linker and non-linker regions among the window sizes from 3 to 20.

In the following MODEL section, we describe the Bayesian model that allows us to compute probabilities of linker state for each residue. We then describe the computation of model parameters using MCMC in the COMPUTATION section. Additional background and details are provided on the supplementary website.

**MODEL**

We assume two hidden states corresponding to the linker and non-linker regions. Let \( Y = (y_1, y_2, \cdots, y_n) \) be the smoothed linker index data of a protein sequence generated by the corresponding hidden state \( S = (s_1, s_2, \cdots, s_n) \). The state transition probability matrix \( P \) given by a two state HMM is \( \{p_{lk}\} = \{p(s_i = k|s_{i-1} = l)\} \), \( l, k \in \{0,1\} \) given

\[
\begin{pmatrix}
  p_{00} & p_{01} \\
  p_{10} & p_{11}
\end{pmatrix} = \begin{pmatrix}
  p_{00} & 1 - p_{00} \\
  1 - p_{11} & p_{11}
\end{pmatrix}
\]

We assume the observed data \( y_i, S \) are independent and have normal distribution. Both the mean and the variance of the observed data are parameterized in terms of the unobserved (hidden) state variable \( s_i \) with a Markov process. If \( s_i = 0 \) then \( y_i \) is from a linker region and if \( s_i = 1 \) then \( y_i \) is from a non-linker region.

\[
s_i = \begin{cases}
  1 & \text{if } y_i \text{ in Non-linker region} \\
  0 & \text{if } y_i \text{ in Linker region}
\end{cases}
\]

By definition of linker index, it is reasonable to give the restriction that the mean linker index of linker region (\( \mu_0 \)) is smaller than the mean linker index of non-linker region (\( \mu_0 + \mu_1 \)), because linker indexes are negative for amino acids that are more prevalent in linker regions.

\[
y_i = \mu_0 + \mu_1 s_i + (1 + \omega s_i)^{1/2} \epsilon_i
\]

Restriction : \( \mu_1 > 0 \)

where the error terms \( \epsilon_i \) are normally distributed with a mean of zero and variance \( \sigma^2 \) (that is \( \epsilon_i \sim N(0, \sigma^2) \)). \( \omega \) denotes the proportionate variance increase when \( s_i = 1 \).

Our objective is to infer the hidden state \( S \), the parameters of model \( \theta = (\mu_0, \mu_1, \sigma^2, \omega) \), and the parameters of transition probabilities \( \eta = (p_{00}, p_{11}) \) given the data \( Y \). We use a Bayesian approach to infer the values of parameters \((S, \theta, \eta)\) from the conditional joint posterior distributions \( P(S, \theta, \eta|Y) \).

The likelihood distribution of data \( Y \) given hidden state \( S \), the parameters of the model \( \theta \) and the parameters of transition probabilities \( \eta \) is

\[
P(Y|S, \theta, \eta) = N((1, S)\mu, \sigma^2 \Sigma)
\]

where the vector of means \( \mu = (\mu_0, \mu_1)' \), \( l = (1, \cdots, 1) \) is a \( n \times 1 \) vector and \( \Sigma = \text{diag}(1 + w_{s_1}, 1 + w_{s_2}, \cdots, 1 + w_{s_n}) \). We can assume \( P(s_i = 1) = 1 \) because the initial state must begin with the non-linker region state in a protein sequence.

The likelihood distribution of the hidden state \( S \) conditioned on the initial state being non-linker is given by

\[
P(S|\theta, \eta) = P(s_1) \prod_{i=2}^{n} P(s_i|s_{i-1})
\]

\[
\propto \beta(n_{00}, n_{01}) \times \beta(n_{11}, n_{10})
\]

where \( n_{ij} \) is the number of observations from state \( i \) to \( j \). Here a random variable \( X \) is said to follow a beta distribution if \( \beta(a,b) \sim X \sim (1 - X)^{a-1} \).

We specify the prior distribution \( P(\theta, \eta) \) in The Prior distributions section below, to complete the conditional joint distribution \( P(S|\theta, \eta|Y) \).

Finally, we calculate the probability state \( k \) for each residue \( i \) in a protein sequence given \( y_i, s_{i-1} = l, \theta \) and \( \eta \). For simplicity, here we show the conditional distribution, suppressing the conditioning on \( \theta \) and \( \eta \).

\[
P(s_i = k|y_i, s_{i-1} = l) = \frac{P(y_i|s_i = k)p_{lk}}{\sum_{j=0}^{1} P(y_i|s_i = j)p_{lj}}
\]

(1)

Once the simulated sample values have been obtained from equation (1), the posterior expectation can be estimated by the sample average, using equation (2).

\[
E(P(s_i = k|y_i, s_{i-1} = l, \theta, \eta)]
\]

\[
= \frac{1}{m} \sum_{t=1}^{m} P(s_t = k|y_t, s_{t-1} = l, \theta^{(t)}, \eta^{(t)})
\]

(2)

where \( t \) denotes the iteration in the MCMC sampler, \( k \in \{0,1\} \) and \( m \) is the number of MCMC samples taken from the posterior distribution after burn-in (early MCMC iterations that reflect the starting value, prior to convergence). We predict the state of an amino acid using the classification variable CV.

\[
CV_i = \begin{cases}
  1 & \text{if } E[P(s_i = k|y_i, s_{i-1} = l)] \leq x \\
  0 & \text{if } E[P(s_i = k|y_i, s_{i-1} = l)] > x
\end{cases}
\]

where \( x \) is the selected cut-off.
The prior distributions
We assign mutually independent prior distributions for $\mu$ and $\sigma^2$. The prior of $\mu$ is assigned to be the conjugated normal distribution and the prior of $\sigma^2$ is the inverse gamma distribution. Here a random variable $X$ is said to follow an Inverse Gamma distribution if $IG(\frac{a}{2}, \frac{b}{2}) \sim (\frac{1}{\xi})^\frac{a}{2} exp(-\frac{\xi}{2} b)$.

$$\mu \propto N\left(\left(\frac{\mu_{0a}}{\mu_{1a}}, 0\right), (0, \xi_{1a})\right)$$

$$\sigma^2 \propto IG\left(\frac{a}{2}, \frac{b}{2}\right)$$

Given hidden state $s_i$, $\omega$ only depends on the observations for $s_i = 1$. We use the expression $\bar{\omega} = (\omega + 1)$ in Albert and Chib (1993) to make $\omega$ represent the proportionate increase in variance when $s_i = 1$. Let the prior distribution of $\omega$ be the truncated inverse gamma distribution $\bar{\omega} \propto IG\left(\frac{a_w}{2}, \frac{2}{b_w}\right) \times I(\bar{\omega} > 1)$.

For the priors for $(p_{00}, p_{11})$, we assign the conjugate beta priors for $(p_{00}, p_{11})$.

$$p(p_{00}, p_{11}) \propto beta(u_{00}, u_{01}) \times beta(u_{11}, u_{10})$$

COMPUTATION
Our challenge in applying the model is to determine the posterior distribution of each parameter. The posterior distribution is not available in explicit form, so we use the MCMC method, Gibbs sampling, to simulate the unknown parameters from the posterior distribution. Details of the computation are provided on the supplementary website.

It is convenient to transform data using $q_i = (1 + \omega s_i)^{1/2}$ so that the transformed data have constant variances instead of variances that depend on state.

$$q_i^{-1} y_i \sim N((q_i^{-1}, q_i^{-1} s_i) \mu, \sigma^2)$$

$$\equiv y_i^* \sim N(w_i^* \mu, \sigma^2)$$

where $I = diag\{1\}$. Define $Y^* = (y_1^*, \cdots, y_n^*)'$ and $W^* = (w_1^*, \cdots, w_n^*)'$.

The full conditional distributions of $\mu$ and $\sigma^2$ are as follows:

$$\mu | \sim N\left(\left(A^{-1}(\sigma^{-2} W^* Y^* + V^{-1} \mu_0), A^{-1}\right) \times I(\mu_1 > 0)$$

$$\sigma^2 | \sim IG\left(\frac{a + n}{2}, b + (Y^* - W^* \mu)'(Y^* - W^* \mu)\right)$$

where $A = (V^{-1} + \sigma^{-2} W^* W^*)^{-1}$, $\mu_0 = (\mu_{0a}, \mu_{1a})'$ and $V = diag(\xi_{0a}, \xi_{1a})$.

The full conditional distribution of $\bar{\omega}$ is the truncated inverse gamma distribution.

$$\bar{\omega} | \sim IG\left(\frac{n_1 + a_w}{2}, \frac{2}{\sum_{i \in J} (y_i - \mu_{0a})^2 + b_w}\right) \times I(\bar{\omega} > 1)$$

where $J = \{i | s_i = 1\}, i = 1, \cdots, n$ and $n_1$ is the number of observations whose state are 1.

The full conditional distributions of $\eta = (p_{00}, p_{11})$ are as follows:

$$p_{00} | s \sim beta(n_{00} + u_{00} - 1, n_{01} + u_{01} - 1)$$

$$p_{11} | s \sim beta(n_{11} + u_{11} - 1, n_{10} + u_{10} - 1)$$

The full conditional distribution of $\{s_i, i = 1, \cdots, n\}$ depends on the state at position $(i-1)$ and $(i+1)$ along a sequence since $s_i$ has a Markov property.

$$P(s_i \mid Y, S^{-i}, \theta, \eta) \propto P(s_i | s_{i-1})P(s_{i+1} | s_i)f(y_i | s_i, \theta, \eta)$$

for $2 \leq i \leq n - 1$

where $S^{-i} = (s_1, \cdots, s_{i-1}, s_{i+1}, \cdots, s_n)'$ and $p(s_1 = 1) = 1$ and $p(s_n = 1) = 1$.

RESULTS
We applied our model to the protein sequence dataset constructed from Pfam-A using linker index $y^4$ as described in DATA. To evaluate the accuracy of the prediction, 5-fold cross validation was applied to the dataset. That is, we divided the dataset into the training dataset and the test dataset randomly with the ratio of 4:1. We trained the model with the training dataset of 642 sequences and tested the trained model with the test dataset of 160 sequences. This procedure was repeated 5 times.

We ran Gibbs sampling with 40,000 iterations and 10,000 burn-in to train the model. The choice of hyper-parameters (parameters of the prior distribution), which are based on the data and the problem at hand, are as follows. We let hyper-parameters for $\mu$ be the sample means of the training dataset for each state and give each a sufficiently large variance of 10. We assume $E(\bar{\omega}) = 1.5$, var($\bar{\omega}$) = 10 and $E(\sigma^2) = 0.1$, var($\sigma^2$) = 10 and fix the hyper-parameters accordingly. We let the hyper-parameters for the $p_{00}$ and $p_{11}$ be $u_{ij} = 1, i, j \in \{0, 1\}$ to have uniform priors. We calculated the probability of linker state, $p(s_i = 0 | y_i)$, for each residue $i$ along a protein sequence.

Figure 1 shows cases with good predictions, in which probabilities in the linker region are much higher than in other regions. However, we need to select a cut-off
value (here, 0.75) to delineate the boundary. Although our method gives high probabilities to the linker region, it also gives high probabilities to other regions that may have similar structure. Figure 2A shows one of these cases. There are two regions with high probability, but there is only one linker region in the protein. The probability of the actual linker region is slightly higher than that of the false positive linker region. Figure 2B shows that sequence termini can have high probabilities.

Figure 3 shows the effect of CV cut-off on $S_n$, $S_p$ and $C$, each averaged over the five tests. Separate curves for each test are provided on the supplementary website. Using a CV cut-off of 0.75, $S_n$ and $S_p$ were each 67%, indicating that we can identify 67% of the linker residues, and that 67% of the residues predicted to be linker are truly linker, respectively. The average Matthew’s correlation coefficient was 65% at the 0.75 CV cut-off.

The test of $S_n$ and $S_p$ described above did not exclude false positives that occurred at N- and C-terminal residues. To test the effect of N- and C-terminal residues on the false positive rate, we recalculated $S_p$, ignoring false positives within the first 20 and last 20 residues of the sequence. The recalculated $S_p$ was 68%, indicating that sequence termini do not contribute significantly to the false positive rate.

There are several other methods for predicting protein linker regions, but it is difficult to compare, because linker definition, type of data required as input and evaluation criteria vary across methods. We compare our method with DomCut (Suyama and Ohara, 2003), because the software is freely available and the authors use a similar linker definition and the same property (linker index) as input; other methods were developed using structural domain/linker definitions and require additional input such as predicted structure or multiple alignments of homologs. DomCut predicts putative linker regions instead of giving specific linker region boundaries, so the DomCut authors evaluated their method in terms of predicted linker regions, not residues (Suyama and Ohara, 2003). This is different than the evaluation of our method described above, which considered predicted linker residues, rather than predicted regions. Therefore, we performed an additional evaluation of our method in order to compare it with DomCut; this time we excluded
the unknown regions in the sequences (the N- and C-termini), and we evaluated linker region predictions, as done in the evaluation of DomCut (Suyama and Ohara, 2003).

A linker region was predicted by our method if the following were satisfied: the length was more than 4 residues, each residue had a high probability (> 0.5), and the maximum probability was greater than 0.8. In DomCut, a linker region is taken to be correctly predicted if there is a trough in the linker region and the minimum linker index value is lower than the cut-off. We tested DomCut using our dataset. The smoothed linker index for the DomCut test was calculated with window size 9. We used a linker index cut-off of -0.08, which we had determined to be optimal for the window size. Sensitivity and specificity of our method were 63.3% and 92.9%, respectively, which appeared to improve the DomCut method (Table 2). Our method correctly predicted 131 linker regions of 141 total predictions, while DomCut correctly predicted 117 linker regions of 113 total predictions. There were 207 linker regions in the dataset.

The low false positive rate of our method in testing linker region prediction compared with the false positive rate in testing linker residue prediction can be attributed to two factors. First, the removal of sequence termini in the region-based test eliminated some false positives. Second, some false positives in the residue-based test were caused by linker boundary extension into domain regions, exemplified by the broad peak in Figure 1B. Overpredicting linker length would not be detected in the region-based test.

<table>
<thead>
<tr>
<th></th>
<th>Comparison</th>
<th>Our Method</th>
<th>DomCut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>63.29</td>
<td>56.52</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>92.91</td>
<td>87.97</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Comparison with DomCut

**DISCUSSION**

We have developed a hidden Markov model for evolutionarily defined protein inter-domain linker/non-linker regions in a protein sequence using the composition differences of amino acids. Results suggest that our method slightly improves an existing method that uses similar biological evidence.

Our choice of dataset has important implications in surmising our method’s ability to predict structurally-defined protein linkers. We used domain definitions provided by Pfam-A, based on evolutionary evidence, rather than structurally defined domain definitions. Our linker/domain dataset is similar to the dataset of Suyama and Ohara (2003), who identified domains using the term 'DOMAIN' in the feature table of the SWISS-PROT database. SWISS-PROT domain annotation is based on the InterPro domain databases, which include Pfam (Apweiler, 2001; Apweiler et al., 2001). We chose to use Pfam-A based domain definitions for SWISS-PROT and SP-TrEMBL proteins, because the resulting non-redundant dataset of multiple domain proteins (802 proteins) was much larger than could be acquired from a structure database (e.g. 101 proteins in Tanaka et al., 2003).

A notorious problem in structural linker prediction has been distinguishing linkers from intra-domain loops. Our method appears to perform well in this regard; however, the relatively low number of false positives may be due to bias in the dataset. Since Pfam-A identifies domains as evolutionarily conserved units, non-conserved intra-domain loops can cause structural domains to be annotated as multiple Pfam-A domains. Thus, some of our Pfam-A defined linkers may actually be loops in structural domains. Conversely, two structural domains that are always found together may be defined by Pfam-A as a single evolutionary domain; some of our false positives may actually be structural linkers. We must test our approach using a structurally defined dataset to fully understand its ability to distinguish structural linkers from intra-domain loops.

We have demonstrated the value of our method in defining linkers for evolutionary domains. Liu and Rost (2003) review methods of computational domain dissection and protein sequence clustering, and suggest that better tools are needed to dissect proteins into domains to cluster them into families. Our method can be used to delineate domains prior to clustering.

We have also demonstrated a HMM approach that considers protein sequence data as continuous variables (linker index) instead of categorical variables (amino acid), and generates probabilistic output. Existing methods that rely on amino acid propensity (Suyama and Ohara, 2003; Miyazaki et al., 2002) do not give probabilistic output. The approach presented here can be extended to other protein sequence/structure problems.

Not only is the composition of the linker region important, but also its length. In general, altering the length of linker regions connecting domains has been shown to affect protein stability, folding rates and domain-domain orientation (van Leeuwen et al., 1997; Robinson and Sauer, 1998). Numerous studies, including that of George and Heringa (2002a) show that the distributions of length of linker and non-linker regions are significantly different. In the future, we can incorporate the informative characteristic of length of linker/non-linker regions by applying a variable duration HMM, which incorporates the specified state duration (i.e. length) distribution.
ACKNOWLEDGEMENTS

We wish to thank the two anonymous referees who provided valuable critiques of this work.

REFERENCES


Robert, C.P. and Mengersen, K. L. (1999) Reparame-
Protein Inter-domain Linker Prediction


