A study of structural properties on profile HMMs

Juliana S Bernardes\textsuperscript{a,}\textsuperscript{*}, Alberto M. R. Dávila\textsuperscript{b}, Vítor Santos Costa\textsuperscript{c}, Gerson Zavurcha\textsuperscript{a}

\textsuperscript{a}COPPE Engenharia de Sistemas e Computação, UFRJ, Rio de Janeiro, \textsuperscript{b}Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil, \textsuperscript{c}CRACS and DCC-FCUP, Universidade do Porto, Portugal

ABSTRACT

Motivation: Profile hidden Markov Models (pHMMs) are a popular and very useful tool in the detection of the remote homologue protein families. Unfortunately, their performance is not always satisfactory when proteins are in the "twilight zone". We present HMMER-STRUCT, a model construction algorithm and tool that tries to improve pHMM performance by using structural information while training pHMMs. As a first step, HMMER-STRUCT constructs a set of pHMMs. Each pHMM is constructed by weighting each residue in an aligned protein according to a specific structural property of the residue. Properties used were primary, secondary and tertiary structures, accessibility and packing. HMMER-STRUCT then prioritizes the results by voting.

Results: We used the SCOP database to perform our experiments. Throughout, we apply leave-one-family-out cross-validation over protein superfamilies. First, we used the MAMMOTH-mult structural aligner to align the training set proteins. Then, we performed two sets of experiments. In a first experiment, we compared structure weighted models against standard pHMMs and against each other. In a second experiment, we compared the voting model against individual pHMMs. We compare method performance through ROC curves and through Precision/Recall curves, and assess significance through the paired two tailed t-test. Our results show significant performance improvements of all structurally weighted models over default HMMER, and a significant improvement in sensitivity of the combined models over both the original model and the structurally weighted models.

Availability: The HMMER-STRUCT tool has been implemented as Perl scripts and as C source code. The structure weighting procedure is available as a patch to the HMMER program. All the test sets, train sets, programs and scripts used in this study are available in http://wiki.biowebdb.org/index.php/Hmmer-struct.

Contact: julianab@cos.ufrj.br

1 INTRODUCTION

One of the major tasks in computational molecular biology is to aid large-scale protein annotation and biological knowledge discovery. Functional characterization of unknown-function proteins is often inferred through sequence similarity search methods, such as BLAST (Altschul et al., 1990) and FASTA (Pearson, 1985). However, when the evolutionary relationship among proteins is distant, methods based on profile hidden Markov models (pHMMs) (Eddy, 1996; Krogh et al., 1994) are known to outperform methods based on sequence similarity search (Gough et al., 2001; Park et al., 1998).

Profile Hidden Markov Models are probabilistic models that are often used to represent groups of homolog sequences. These models have been a key tool in protein annotation, and are highly effective for scoring similar sequences. Unfortunately, the performance of pHMMs degrades for sequences in the twilight zone, that is, for homologue sequences with low identity (below 30%). This limitation has motivated a number of different approaches to increase pHMM performance. Proposals include new scoring functions, new null models (Karplus et al., 2005) and prior probability (Brown et al., 1993). Researchers have also combined other information with pHMMs: T-HMM (Qian et al., 2004) uses phylogenetic information; HMM-STR (Bystroff et al., 2000), combines pHMMs and support vector machines (Schölkopf et al., 1999).

The observation that homologue proteins tend to preserve structure suggests that structural information should be extremely relevant in detecting homologues. In fact, it has been shown that pHMMs trained with multiple sequence alignments based on proteins structural alignment can have better performance than pHMMs based on state-of-the art aligners that apply primary sequence information only, when remote homology detections are assessed (Bernardes et al., 2007). In this vein, researchers have proposed special alphabets to represent structural elements in pHMMs (Goyon et al., 2004), or modifying pHMM structure to add protein three-dimensional information (Alexandrov et al., 2004). Although such methods are more powerful than pHMMs, arguably they are computationally more expensive both in training and in classification, and to the best of our knowledge have not become widely used.

We present a new method to apply structural information in protein classification. In contrast to the previous approaches, our method relies on pHMMs. Our main contribution is a residue weighting-algorithm that incorporates protein structural information into pHMMs. Further, we apply different structural properties to train a library of 5 pHMMs from a homologue protein set. The properties we consider are primary, secondary, and tertiary structure, also used in previous methods. We also apply two properties that, to the best of our knowledge, have not been used in this task before, but that are often important in this domain: solvent accessibility and residue packing. The classification of a unknown-function protein is then obtained by combining the classification from the library of pHMMs. The main advantage of our method is that structural information is only used to train the pHMMs. Notice that scoring is still performed using sequence data, as opposed to (Alexandrov et al., 2004). Our method was implemented by extending the HMMER package and experimental evaluation using the SCOP database showed significant improvement over HMMER.

\textsuperscript{*}to whom correspondence should be addressed
2 METHODS

In our experiments, a protein homologue set is aligned by utilizing the MAMMOTH-mult (Attwood et al., 2005) aligner. MAMMOTH produces two outputs: one represents the multiple sequence alignment (based on spatial coordinates similarity) and the other the structural alignment. These outputs are used to build weight matrices $M_s$, which represent residue structural weights for each protein. These matrices were used on pHMM training stage. The section 2.2 will give more details on the building of structural weights for each protein. These matrices were used both in the homologue core structure, which is then used to construct pHMM3D. The section 2.2 will give more details on the building of structural weights for each protein. These matrices were used both in the homologue core structure, which is then used to construct pHMM3D.

Fig. 1. First, a homologue protein set is aligned by using MAMMOTH-mult aligner. The aligner produces a multiple sequence alignment and a structural alignment. The multiple sequence alignment is used to build a conventional pHMM using HMMER package, pHMM1D. Aligned sequences are fed to the joy tool. The joy output is used to construct weight matrices, which are then used to build secondary pHMM2D, accessibility pHMMacc, and packing pHMMoi models. Finally, the structural alignment is used to find the homologue core structure, which is then used to construct pHMM3D.

2.1 Profile HMMs

Profile HMMs represent conserved regions in sequences as sequences of match (M) states. Inserted material is represented as insert states (I), and deleted regions as delete states (D). The parameters of PHMMs are probabilities of two events: a transition probability from a state to another state, and a probability that a specific state will emit a specific residue (say, a specific amino-acid when comparing proteins), called emission probability. Obviously, only match and insert states generate characters and have an emission probability distribution; delete states are quiet. In the case of proteins, emission distributions have 20 entries, one per amino-acid.

Possible transitions define the structure of the pHMM. Systems such as SAM (Hughey et al., 1996) allow transitions between all types of states, totaling 3 transitions per state, hence 9 per node. On the other hand, the HMMER system relies on the Plan7 model (Eddy, 1998), which disallows $I \rightarrow D$ and $D \rightarrow I$ transitions.

Emission probabilities are calculated by the equation 1, where $c_j(\sigma)$ is the observed frequency of residue $\sigma$ in $j$ column of the alignment, and $\alpha_j(\sigma)$ represents the pseudo counts of residue $\sigma$, which are obtained from Dirichlet mixtures, as seen in (Brown et al., 1993).

$$e_j(\sigma) = \frac{c_j(\sigma) + \alpha_j(\sigma)}{\sum_k c_j(k) + \alpha_j(k)} \quad (1)$$

In the same way, the transition probability can be found through the equation 2, where $c_{kl}(\sigma)$ is the observed frequency of transitions between state $k$ and state $l$, where $k, l \in \{M, I, D\}$, and $\alpha_{kl}$ represents the pseudo counts of transition between $k$ and $l$.

$$t_{kl} = \frac{c_{kl} + \alpha_{kl}}{\sum_j c_{kj} + \alpha_{kj}} \quad (2)$$

2.2 Sequence Weighting

One problem in representing families of sequences is that often sets of very similar sequences may be over-represented in the training sequences, introducing bias. Therefore, sequence-weighting methods were introduced to compensate for over-representation among multiply aligned sequences. In general, very similar sequence receives lower weights and divergent sequence higher weight. Sequence weighting was applied to the construction of position-specific score matrix (PSSM) (Gribskov et al., 1987), and is fundamental to the performance of profile HMMs. In the latter case, the default sequence weighting method used by HMMER package is a high quality algorithm based on phylogenetic trees (Gerstein et al., 1994).

Let $A$ be an generic aligned alignment used to train a pHMM. Suppose, $A$ with $N$ sequences and length $L$. Then, we can represent $A$ alignment weights as a matrix $W$, such that $w_{ij}$ represents the weight of an amino-acid of protein $i$ in the $j$th alignment position, as shown in the equation 3. Basically, a sequence-weighting method for pHMMs attributes equal weights to all residues in the protein, that is, $w_{ij} = w_{ik}$ for $\forall j, k \leq L$.

$$W = \begin{pmatrix} w_{11} & \cdots & w_{1L} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{NL} \end{pmatrix} \quad (3)$$

In the spirit of PSSMs, we propose to reinforce residues that correspond to preserved regions in the protein. Our motivation is that when homologue proteins are structurally aligned, spatial overlapping of an atom set occurs. This set is called the invariant core or core structure, and can be used to characterize homologue proteins. We argue that the residues in the core structure should carry more weight rather than the residues outside the core. Thus, we propose sequence-weighting method that gives different weight to each residue in the same protein, based on structural relevance. We will represent such "structural" weights by a matrix $M_s$, where each residue of the same protein has a different weight.

$$M_s = \begin{pmatrix} m_{11} & \cdots & m_{1L} \\ \vdots & \ddots & \vdots \\ m_{N1} & \cdots & m_{NL} \end{pmatrix} \quad (4)$$

As mentioned before, the default sequence-weighting method used by HMMER package is a high quality algorithm. Therefore, we combine both the default HMMER’s $M$ matrix in (3) and the of $M_s$ structural matrix in (4), as shown in (5).
\[
\mathbf{M}_a = \mathbf{MM}_s^F = \begin{pmatrix}
  w_{11}M_{11} & \cdots & w_{1L}M_{1L} \\
  \vdots & \ddots & \vdots \\
  w_{N1}M_{N1} & \cdots & w_{NL}M_{NL}
\end{pmatrix}
\]  

However, introducing weights affect the computation of the observed frequencies. More precisely, the observed frequency \( c_j(\sigma) \) shown in 1 is now found through the equation 6, where \( s_{ij} = w_{ij}m_{ij} \) is structural weight of residue \( \sigma \), according to \( M_s \) matrix.

\[
c_j(\sigma) = \sum_i^N \frac{f(\sigma) \cdot f(\sigma)}{f(\sigma) \cdot f(\sigma)} = \begin{cases} 
  s_{ij}, & \text{if } \sigma \text{ is the amino-acid in position } ij \\
  0, & \text{otherwise}
\end{cases}
\]

In the same way, we apply the equations 7 to determine \( c_{kl} \) shown in 2. If the \( k \) and \( l \) states are either \( M \) or \( I \) states, \( c_{kl} \) can be calculated through the arithmetic mean of \( m_{ik} \) and \( m_{il} \). If at least one state is a \( D \) state, \( c_{kl} \) is either \( m_{ik} \), if \( l \in \{D\} \), or \( m_{il} \), if \( k \in \{D\} \). Last, if both are \( D \) states, \( c_{kl} \) is 1.

\[
c_{kl} = \sum_i^N \frac{f_{kl} \cdot f_{kl}}{f_{kl} \cdot f_{kl}} = \begin{cases} 
  \frac{s_{ik} + s_{il}}{2}, & \text{se } k,l \in \{M, I\} \\
  s_{ik}, & \text{se } l \in \{D\} \text{ e } k \notin \{D\} \\
  s_{il}, & \text{se } k \in \{D\} \text{ e } l \notin \{D\} \\
  1, & \text{se } k,l \in \{D\}
\end{cases}
\]

2.3 The \( M_s \) structural weight matrices

As explained above, our algorithm considers a number of different sources of structural information. Next, we approach how this information was obtained and used to build \( M_s \) matrix.

2.3.1 Secondary structural elements Secondary structure is often conserved among homologue proteins. Indeed, motifs (Branden et al., 1991), consensus sequences in homologue proteins, usually include a combination of well conserved secondary structure elements (Chakrabarti et al., 2004).

In order to build a \( M_s \) matrix based on secondary structure elements we need to identify secondary structure elements in the original sequences. This is possible because we assume we have full structural data for the training sequences. In this work, we chose to utilize the SSTRUCT program, part of the widely used JOY package (Mizuguchi et al., 1998), to extract secondary elements from the PDB files. SSTRUCT output is a character sequence, such that the characters \{L=loop, H=helix, C=sheet\} match a secondary structure element against a residue, as shown in figure 2. Following Deane’s work on the relative frequency of conserved regions (Deane et al., 2003), we mapped each SSTRUCT element as follows: \( L \rightarrow 1 \), \( H \rightarrow 2 \), and \( C \rightarrow 4 \). Our mapping thus favours conservation in sheets, and gives default weight to loops. Although the active site of proteins can be found in loops, these regions often contain indel segments. Figure 2 shows an example of structural weight attributions for proteins in a partial alignment.

2.3.2 Solvent Inaccessibility The hydrophobic interactions of non-polar side chains in amino-acids are believed to contribute significantly to the stability of the tertiary structures in proteins. Hydrophobic amino-acids will tend to cluster together, not as a result of attraction, but as a result of their repulsion by the hydrogen bond water network in which the protein is dissolved. Therefore, these amino-acids will preferentially be located away from the surface of the molecule. Since they form the core of protein, they tend to be more conserved and are, thus, more useful for identifying remote evolutionary relationships.

We have utilized the PSA (Lee et al., 1971) program to provide solvent inaccessibility information. PSA is part of the JOY package. The \( M_s \) matrix was built giving weight 3 for inaccessible residues and weight one to the others. The weights are based on (Chakrabarti et al., 2004), which demonstrated empirically that inaccessible amino-acids are three times more conserved than accessible amino-acids. The \( M_s \) matrix represents structural weights that were used to build the model \( pHMM\)acc, as shown in figure 1.

2.3.3 Packing density The tertiary structure of proteins stems from a very large number of atomic interactions. In regions where the interactions are stronger residues tend to be packed together. It is well known that densely packed regions tend to be preserved, and hence that amino-acids belonging to those regions are usually more conserved than other amino-acids. TJ Ooi created a measure, called the Ooi Number (Nishikawa et al., 1986), that estimates the amino-acid packing density. Essentially, the Ooi number counts for a residue counts the number of neighboring C-\( \alpha \) atoms within a radius of 14Å of the given residues own C-\( \alpha \). Although crude, this measure does give a good impression of which parts of the structure are buried and which are exposed on the surface.

We again use the JOY package to obtain the Ooi number and estimate packing density. Figure 3 shows a stretch of JOY output, in which the numbers represent the Ooi measure for the Dehaloperoxidase protein in the Globins family (16wc PDB code). We used these numbers to build the structural weight matrix \( M_s \). The structural weights were then used to build the model \( pHMM\)Ooi, as shown in figure 1.

2.3.4 Homologous Core Structure Structural similarity among proteins can provide valuable insights into their functionality. One way to
provide structural similarities is through three-dimensional alignment of proteins, also called structural alignment. The goal is to align two or more proteins by trying to overlap the three-dimensional coordinates of their atoms. When multiple homologue proteins are structurally aligned, we tend to observe that there is a subset of coordinates whose spatial locations are better conserved across structural alignment. This subset is called the homolog core structure (HCS) (Matsuo et al., 1999). According to the result reported by Gerstein et al. (1995), HCS can be utilized to detect homologue proteins.

Our goal was to estimate the HCS of a set of protein. As a first approximation, we propose a method to extract it from structural alignment by calculating how much aligned residues from different proteins tend to be close together. Following MAMMOTH, we represent residues through the coordinates of their C-α atoms. In other words, we assume that closeness between C-α atoms will approximate overlapping among amino-acids. To find out how much amino-acids are close together, we utilize the Euclidian distance measure, as shown in the equation 8. It represents the shortest distance between two points in the space.

\[
d_{a,b} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}
\] (8)

The degree of overlap between aligned residues in the structural alignment was calculated through the relative distance \(d_{ij}\), equation 9. This distance can be found through the average distance among the amino-acid in the position \(ij\) and other amino-acids in the \(j\) column of alignment.

\[
d_{ij} = \frac{\sum_{b=1}^{n-1} d_{(i,b),(i,b+1)}}{n-1}
\] (9)

Finally, the relative distance was normalized according to 10, and it was used to determine the degree of overlap of each residue. These measures were normalized by using the equation 10, where \(d_{\text{min}}\) is the minimal distance and \(O_{\text{max}}\) is the maximal Ooi measure for protein \(i\).

\[
m_{ij} = \frac{d_{\text{min}} \times O_{\text{max}}}{d_{ij}}
\] (10)

After this step, we built the \(M_s\) matrix, where each \(m_{ij}\) matrix element corresponds to the relative distance of amino-acids \(ij\) in the structural alignment. This matrix represents structural weights that were used to build the model \(p\text{HMM3D}\), shown in the figure 1.

### 2.4 Library of structural models

In a second step, we join the models built from these matrices to form a library of structural models aiming at building a single model to represent the structural patterns under different aspects. We used the \texttt{hmmpfam} HMMER tool to combine the models together. Library of models have been used in a number of studies, such as (Bateman et al., 2004; Haft et al., 2003; Gough et al., 2001), and they are known to achieve better results than those achieved by single models.

### 2.5 Test Procedure

The main concern of our study is to build pHMMs that can be helpful in remote homology detection. Therefore, our experiments considered proteins with identity below 30%. To do so, we used the SCOP database (Andreeva et al., 2004), and more specifically ASTRAL SCOP version 1.67 PDB40 (with 6600 protein sequences). ASTRAL SCOP is particularly interesting for our study because it describes structural and evolutionary relationships among proteins, such that none of the sequences in ASTRAL SCOP present > 40% sequence identity. Thus, it is an excellent dataset to evaluate the performance of remote homology detection methods and has been widely used to reach this goal (Espadaler et al., 2005; Wistrand et al., 2005; Hou et al., 2004; Alexandrov et al., 2004).

SCOP classifies all protein domains of known structure into a hierarchy with four levels: class, fold, super family and family. In our study, we work at the super family level, which gathers families in such a way that a common evolutionary origin is not obvious from sequence identity, but probably from an analysis of structure and from functional features. We believe that this level better represents remote homolog.

Moreover, we used cross-validation (Mitchell, 1997) to compare the different approaches. First, we divided SCOP database by super family level. Next, from ASTRAL PDB40, we chose those super families containing at least three families and at least 20 sequences. We eventually tested 39 super families, as listed in Table 1. This whittled down the number of sequences we used to model building to 1137. Third, we implemented leave-one-family-out cross-validation. For any super family \(x\) having \(n\) families, we built \(n\) profiles so that each profile \(P\) was built from the sequences in the remaining \(n - 1\) families. Thus, the \(n - 1\) sequences form the training set for profile \(P\). The test set for profile \(P\) will be the remaining sequences (test positives) plus all other database sequences (test negatives).

### Table 1. Superfamily SCOP-Ids

| Superfamily SCOP-Ids | 
|----------------------|
| a.1.1. | a.138.1. |
| a.25.1. | a.26.1. |
| a.3.1. | a.39.1. |
| a.4.1. | b.121.4. |
| b.18.1. | b.29.1. |
| b.36.1. | b.47.1. |
| b.55.1. | b.60.1. |
| b.6.1. | b.71.1. |
| b.82.1. | c.1.10. |
| c.23.1. | c.26.1. |
| c.36.1. | c.52.1. |
| c.55.1. | c.55.3. |
| c.67.1. | d.108.1. |
| d.14.1. | d.144.1. |
| d.15.1. | d.153.1. |
| d.169.1. | d.3.1. |
| d.58.7. | d.92.1. |
| g.3.11. | g.3.6. |
| g.3.7. | g.37.1. |
| g.39.1. | |

SCOP Super families used in our experiments. We only considered super families with at least 20 proteins and three or more families.

In order to assess HMMER-STRUCT performance, we used the HMMER package. We did not compare with SAM (Hughey et al., 1996) package. First, because our goal was to evaluate whether structural properties can improve pHMMs, not to compare the two packages, and second, because a related previous study on the same dataset actually showed HMMER outperforming SAM (Bernardes et al., 2007). The same study also indicated better results on the “twilight zone” using structural alignment tools, such as MAMMOTH-mult and 3DCOFFEE. We used MAMMOTH in this study.

Results were graphically analyzed by building ROC and Precision/Recall curves. ROC curves are a common measure of performance that is very used in bioinformatics application. They are based on the relation of the false positives (non homologue proteins) and of true positives (homologue proteins), and are obtained by varying a parameter that affect these relationships. We further present Precision/Recall curves, as they give a good perspective on true positives, false positives and false negatives hits. In both cases, the bigger the area under the curve (AUC), the more efficient the analyzed tool is. In both cases we used the minimal \(e\)-value required to accept a match as the parameter used to build both curves. We ranged \(e\)-values between \(10^{-50}\) and 10. Finally, we used the paired two tailed t-test to assess significance, and assumed that results with \(p \leq 0.05\) (i.e. 95% of confidence) are significant.

### 3 RESULTS

As a first step, we build a model for each structural property and evaluate it according to the methodology described in the Methods section. The ROC curves are presented in figure 4 and the Precision/Recall curves in figure 5. Both figures show all models, that is, \(p\text{HMM2D}\) (secondary structural model), \(p\text{HMMOi}\) (Ooi measure model), \(p\text{HMMAcc}\) (inaccessibility model) and \(p\text{HMM3D}\) (three-dimensional structure model) outperforming the HMMER model. Table 2 shows the paired two tailed t-test between each model. All
models built from structural properties perform significantly when compared to HMMER. Only, the pHMMD and pHMMAcc results are not significant in relation to each other.

![ROC Curve](image1.png)

**Fig. 4.** Performance of each model in HMMER-STRUCT tool, for MAMMOTH aligner, as measured by ROC Curves

![Precision/Recall Curve](image2.png)

**Fig. 5.** Performance of each model in HMMER-STRUCT tool, for MAMMOTH aligner, as measured by Precision/Recall Curves

| Model     | pHMM2D | pHMM3D | pHMMAcc | pHMMOi |
|-----------|--------|--------|---------|--------|
| pHMMD     | 0.0074 | 0.04   | 0.01    | 0.01   |
| pHMMAcc   | 0.0387 | 0.01   | 0.05    |        |
| pHMM3D    | 0.01   | 0.01   |         |        |
| pHMM2D    | 0.0066 |        |         |        |

**Table 2.** HMMER-STRUCT paired t-test

Paired two tailed t-test when comparing performance of each HMMER-STRUCT model all against all.

Next, we compare the performance of the model library with respect to the initial HMMER model. To do so, we joined the five models, one for each structural property, and scored the test sequences using hmmpfam. Figure 6 shows the ROC curve for the results. Figure 7 shows graphically the results through Precision/Recall curves. Both figures show HMMER-STRUCT outperforming HMMER. Table 3 displays significance results. The difference between HMMER-STRUCT and HMMER results are statistically significant according to paired two tailed t-test. The two tailed t-test also indicate significant differences between HMMER-STRUCT and each HMMER-STRUCT component, i.e, HMMER, pHMM2D, pHMM3D, pHMMAcc and pHMMOi.

![ROC Curve](image3.png)

**Fig. 6.** HMMER-STRUCT Performance for MAMMOTH aligner, as measured by ROC Curves

![Precision/Recall Curve](image4.png)

**Fig. 7.** HMMER-STRUCT Performance for MAMMOTH aligner, as measured by Precision/Recall Curves

4 DISCUSSION

The accuracy of homology detection methods is essential for the problem of inferring the function of unknown-function proteins.
However, improving accuracy becomes hard when similarity between sequences is low. We proposed a method to improve pHAMMs sensitivity by adding structural properties in the model building stage. We showed that the pHAMMs trained according to this method are more sensitive than pHAMMs trained from multiple sequence alignments, even if the alignment itself relied on structural properties. Our experiments demonstrated best performance for pHMM2D, that used secondary structural properties, and for pHMMOi, that used packing density residues. Both pHMMs present similar performance. We believe that the good results obtained with the pHMMOi model can be attributed to the fact that tight packing is important for protein stability, and follow well-known results that indicate that amino-acids located in the core protein are more conserved than amino-acids located in other sites (Privalov, 2000). In the same way, the pHMM2D model achieve good performance as secondary structure elements are responsible for maintaining the form in homologue proteins. These elements form motifs and domains, which are related with protein function. Conserved sites may point to functionally and structurally important regions. These observations may explain the higher performance of models based on packing residues and on secondary structural properties.

The pHMMAcc models, based on amino-acid inaccessibility, and the pHMM3D models, based on three-dimensional coordinates, did not perform as well. The pHMMAcc models did not achieve statistical significance results, when they were compared with HMMER. On the other hand, we observe that the inaccessibility property can be explained by hydrophobic effects, as are the amino-acids with hydrophobic side-chain that go toward the core protein by forming packages. Therefore hydrophobicity was represented in the pHMMOi model, that achieved good performance. Our results suggest the difference between models stems from the pHMMOi models to be more accurate and precise than what is used when building pHMMAcc.

However, we believe the inaccessibility property is already represented appropriately by pHMMOi model. Since amino-acids with high packing density already are inaccessible. Therefore, pHMMOi outperformed the pHMMAcc, as pHMMOi has more information than pHMMAcc.

The chief contribution of our method was achieved when all the models work together. The combined models performed significantly better than any single model. We believe that this results from the fact that each trained pHMMs represents a different structural property. Therefore, combining the models increases sensitivity by exploring the different structural properties.

Our method shows that structural information can be added during the training phase of pHMM to improve sensitivity, without much changes to the usage of pHMM methodology, and applied to recently discovered proteins for which there is little structural information.

## 5 CONCLUSION

The increasing number of studies involving pHMMs and the use of structural information has been quite remarkable (Hou et al., 2004; Alexandrov et al., 2004; Bystroff et al., 2000). Most of these approaches build structural models based on three-dimensional coordinates. In contrast, we present a novel methodology to train pHAMMs based on structural alignment and other structural properties using a set of homologue protein sequences. Our method builds five models from an aligned homologue sequence set. Each model represents a different structural property, and the union of the models represent the structural context of aligned proteins. The properties used were primary, secondary and tertiary structures, accessibility and packing residue. Note that previous attempts have already used secondary and tertiary structural properties to train pHMM, though in quite a different way. However, accessibility and packing residue properties were used for the first time in pHMM training, with good results in the latter case.

In order, to build each model, we developed a novel sequence-weighting algorithm based on structural weights that are attributed for each amino-acid. Traditional weighting-algorithm works gives the same weight for every residue in the protein. Instead, we propose a method that gives a different weight to each amino-acid into a protein, according to structural properties that suggest it may be in a conserved region. Our results relied on prior work (Chakrabarti et al., 2004; Deane et al., 2003; Nishikawa et al., 1986) that suggested interesting properties and estimated their weight.

Nowadays, the most popular approach to discovering the function of a newly found protein is through sequence similarity search. In fact, it is well known that structure is more conserved than sequence, and thus structural similarity can suggest function similarity. On the other hand, structural data is sparse and are usually not available for proteins with unknown function. Therefore, it is very important that methods that uses structural properties to build models will not need to rely on structural information for a new protein. Our method makes use of structural properties only at the model building stage, but not at scoring.

Our results show that the use of structural properties can improve the sensitivity of remote homology methods. Moreover, the combination of different model (one for each property) outperforms the use of individual properties. A number of future research directions present themselves. It will be interesting to include more models, such as that based on bond-hydrogen properties. Also, it will be interesting to apply our methodology to other remote homology tools, such as that based on bond-hydrogen properties. Ultimately, we believe that our work is a step in the major challenge of finding the set of structural properties or features that represent precisely membership of a super family.

## ACKNOWLEDGEMENT

We are grateful to CNPq for financial support.

| Table 3. HMMER-STRUCT paired t-test |
|--------------------------------------|
| HMMER | pHMMAcc | pHMMOi | pHMM2D | pHMM3D |
| $10^{-4}$ | $10^{-4}$ | $10^{-5}$ | $10^{-5}$ | $10^{-4}$ |

Paired two tailed t-test when comparing performance of each HMMER-STRUCT component with the combined model.
