A comparative study of HMM classifier on protein sequence classification

Wajdi DHIFLI1,2,3, Rabie SAIDI1,2,3, Jalel AKAICHI4, Engelbert MEPHU NGUIFO1,2

1 Clermont Université, Université Blaise Pascal, LIMOS, BP 10418, F-63000 Clermont-Ferrand, France
2 CNRS, UMR 6158, LIMOS, F-63173 Aubière, France
3 Université de Jendouba, faculté des sciences juridiques, économiques et de gestion, Jendouba, Tunisie
4 Université de Tunis, Institut Supérieur de Gestion, Tunis, Tunisie

Background and aims

Background

Generally speaking, proteins do everything in the living cells. Hence, the task of protein classification is very important in bioinformatics. Several techniques were proposed in literature dealing with this task. In [1] authors combined different encoding methods with well-known machine learning classifiers then compared the results with those of the alignment based classification using Blast [2]. Five datasets with dissimilar characteristics were used.

Aims

- Using hidden markov models (HMM) [3] to classify biological sequences.
- Comparing HMM’s results with other well known classifiers.
- Evaluating the effect of the dataset characteristics variation on HMM accuracy.

Methods

Hidden Markov models (HMM)

![HMM Diagram](Image 1350x1544 to 2213x1658)

Hidden states

Observations

Emitted sequence

Fig 1. Example of hidden Markov model.

In the case of protein sequences [4], “observations” represent the emitted amino acids sequence while “match”, “insert” and “delete” represents the hidden states. Arrows represent both transitions and emissions probabilities.

Experimental setup

Experiments were run under Linux using 2Gb Intel core 2 duo processor and 4Gb RAM DDR3. We used the leave one out (LOO) technique for evaluation. Blast and Weka classifiers were used with default values.

Experimental results

In DS1, DS2 and DS5 all classifiers scored very high since these datasets presented good inter-class and intra-class similarity and balanced classes. DS3 was a real challenge since it comprised seven unbalanced classes with low intra-class similarity. Therefore, HMM did not score as good as in the others, since the generated profiles were very poor with low discrimination power. However, in DS4 the intra-class similarity was low too but class sizes were close, though no full accuracy was reached however all classifiers performed well.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Blast</th>
<th>HMM</th>
<th>C4.5</th>
<th>SVM</th>
<th>NB</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>100</td>
<td>100</td>
<td>96.7</td>
<td>96.7</td>
<td>90</td>
<td>78.3</td>
</tr>
<tr>
<td>DS2</td>
<td>100</td>
<td>99.21</td>
<td>99.8</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>DS3</td>
<td>69.60</td>
<td>28.73</td>
<td>79.2</td>
<td>78.94</td>
<td>59.4</td>
<td>77</td>
</tr>
<tr>
<td>DS4</td>
<td>78.57</td>
<td>70</td>
<td>82.5</td>
<td>87.5</td>
<td>95</td>
<td>80</td>
</tr>
<tr>
<td>DS5</td>
<td>78.3</td>
<td>82.67</td>
<td>75.5</td>
<td>84.1</td>
<td>85.9</td>
<td>80.5</td>
</tr>
</tbody>
</table>

Table 1. Experimental results.

Conclusion

This study proved that using HMM as classifier represents a competitive approach. However, it may fail in some cases especially with datasets having unbalanced classes, low intra-class and/or high inter-class identity. This yield generating poor profiles and decrease the discrimination power of the classifier. Therefore, instead of blaming the classifier, studying the used data characteristics before choosing the classifier would be a wise decision.

Acknowledgment

A fellowship was granted to the first author of this work from the “higher education and scientific research ministry” in Tunisia.

References