Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Manuscript Category

OXFORD

Protein Classification

Predicting Protein Origin from Sequence Data

Eric Hambro [1,*]

[1] Department, Institution, City, Post Code, Country

* To whom correspondence should be addressed.

Abstract

Motivation: Deriving protein function from sequence level data can provide. Using features derived from the sequence and a random forests model was trained to classify protein sequences according to their location of use. Using a validation set of blah, an accuracy prediction accuracy of blah was reported, along with an F1 score of blah.
Results: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Availability: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Contact: eric.hambro.17@ucl.ac.uk

1 Introduction

Although sequence data of genes has proliferated since the advent of **technique**, the classification of protein function has lagged behind significantly. **Statistic**. This is partly down to the fact that protein function is a complex labelling task, given the function of proteins may vary under different conditions. In cases where no homologues are present for a gene, knowing the subcellular location of the the protein gives indicator as to the proteins function. Furthermore discovering subcellular location has proven to be vital information in the research of new drugs and vaccinations. As such, automated means of predicting subcelular localization are utmost urgency in todays drug search enbironment.

Since as early as **blah**, machine learning has been used to predict automatically the subcellular localization of a protein, in an attempt to reduce the cost and time taken to process the proliferating protein sequences. Traditional methods such as **experiment**, which were laborious and time intensive, have been replaced in parts by papers, which are able to make predictions at a high degree of confidence *reference* with only sequence data.

Some these techniques include **machine learning technique**, which does a **xyz**, and **ml technique 2** which does **this**.

In this particular experiment we are presented with a training data set *X* proteins, each corresponding to one of four subcellular locations, and we train and evaluated a Random Forests classifier on these data, finally making predictions of locations for *X* unknown proteins.

Section 2 of this report contains a description of the method used, and Section 3 presents our results. A discussion of these results is then presented in Section 4, along with analyses of the features of the model, a comparison to another model, and our motivation in choosing this method. Finally, I present a comparison with existing work in the field and further work that could be done, ending with a conclusion.

2 Method

2.1 Preprocessing

Data was obtained from four files containing protein sequences in Fasta format. Each file contained proteins according to its type of subcellular localization: cytosolic, secreted, nuclear and mitochondrial. The number of proteins in each file is presented in Table 1. These data files contained no homologues, and therefore could be assimilated, shuffled and split into a training and validation set randomly.

The data was then fed through a pipeline to extract a number of features from the sequence. Here the open source 'BioPython' *reference* module was used, which allowed us to augment sequence data with important characteristics of the amino acids, such as hydrophobicity or molecular weight. These features were :

- Feature 1 What is each one. a line or two.
- Feature 2
- Feature 1
- Feature 1
- Feature 1

.

The numerical features were then concatenated to form a vector, and divided such that 30% of the data formed a validation set, with the remainder forming a test set. An anlysis of the features themselves is present in section 4.

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Table 1. Sample Data Set By Protein Class

Cytosolic	Nucleic	Secreted	Mitochondrial
row4	row4	row4	row4

2.2 Classifier Training

A Random Forest Classifier was used in the training of the model, which was done with 70% of the training data. The classifier itself was implemented using the open source 'scikit-learn' package, and was instantiated with the a **hyperparameters of X, Y, Z**. These hyperparameters were all chosen in advance using k-fold cross validation, and a grid search on hyper parameters.

2.3 Cross Validation & Hyperparameter Tuning

In order to assess both the strength of model under certain hyperparameters, and the strength of the predictive power of certain features, a 5-fold cross validation score *(what score?)* of the model was used, each time with a 70%, 30% train-validation split. In order to search the space of hyperparameters, a grid search was established, visible in plot *plt the cross valid hyper param search*.

2.4 Ablation Study

An ablation study was performed, by removing one feature at a time and testing the 5-fold cross validation score of the model in predicting these data. The results are presented *fig xyz*

2.5 Gradient Boosting Comparison

The training set of features was then

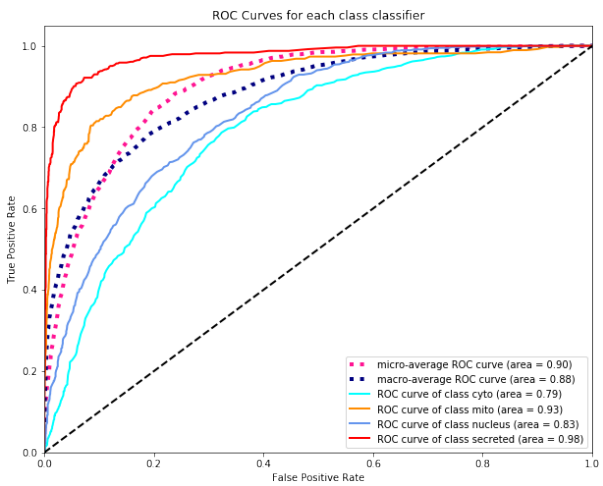


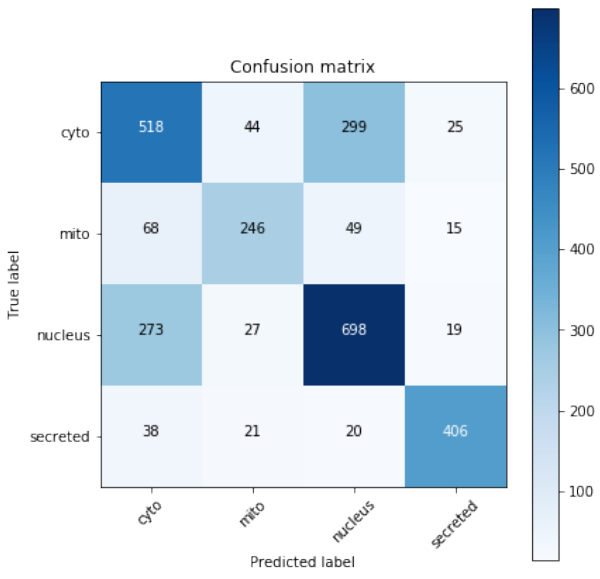
Table 2. Predictions of Test Proteins

Protein ID	Subcellular Location	Probabilty
row1	row1	row1
row2	row2	row2
row3	row3	row3
row4	row4	row4

This is a footnote

3 Results

yip



- Results of accuracy - Confusion matrix - ablation study - results compared to xgboost

Obtained data from X files, and mixed, shuffle and split into a train and validation split. Since the proteins are all non homolgoues, no need to arrange by family and could split anyway.

Then took these files and processed them into feature vectors, using the bipython module. These extracted, x features from the sequence.

Then trained a random forest classifier with X trees, and also tested a XGBboost classifier.

Cross validated grid search done to find best parameters for trees Ablation study used to investigate salient features.

Trained a random forest with X trees.

What did you do? - extract features - train -test split - homologies - cross validate - compared with a model (gradient boosting?) - ablation study

- Results of accuracy - Confusion matrix - ablation study - results compared to xgboost

4 Discussion

- Examine the features - Examine the ablation study - Examine comparison to XGBoost - Compare with existing literature and approaches

5 Further Work

- suggestions: more data? better features/

6 Conclusion

7 Theoretical Background

What about ???

end of paper

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 1 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. (Bag *et al.*, 2001) wants to know about …… text follows.

$$\sum x + y = Z \tag{1}$$

8 Approach

Equation (1) Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text .....

9 Methods

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text

- for bulleted list, use itemize
- for bulleted list, use itemize
- for bulleted list, use itemize

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

9.1 This is subheading

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text. Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might

want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

9.1.1 This is subsubheading

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text textText Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the

Table 3. This is table caption

head1	head2	head3	head4
row1	row1	row1	row1
row2	row2	row2	row2
row3	row3	row3	row3
row4	row4	row4	row4

This is a footnote

Fig. 1. Caption, caption.

above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text

9.2 Test1

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text

10 Discussion

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text.

Table 3 shows that Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text.

11 Conclusion

(Table 3) Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text

1. this is item, use enumerate
2. this is item, use enumerate
3. this is item, use enumerate

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above

method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text

Acknowledgements

Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text

Funding

This work has been supported by the... Text Text Text Text.

References

Bofelli,F., Name2, Name3 (2003) Article title, *Journal Name*, **199**, 133-154.

Bag,M., Name2, Name3 (2001) Article title, *Journal Name*, **99**, 33-54.

Yoo,M.S. *et al.* (2003) Oxidative stress regulated genes in nigral dopaminergic neurnol cell: correlation with the known pathology in Parkinson's disease. *Brain Res. Mol. Brain Res.*, **110**(Suppl. 1), 76–84.

Lehmann,E.L. (1986) Chapter title. *Book Title*. Vol. 1, 2nd edn. Springer-Verlag, New York.

Crenshaw, B.,III, and Jones, W.B.,Jr (2003) The future of clinical cancer management: one tumor, one chip. *Bioinformatics*, doi:10.1093/bioinformatics/btn000.

Auhtor,A.B. *et al.* (2000) Chapter title. In Smith, A.C. (ed.), *Book Title*, 2nd edn. Publisher, Location, Vol. 1, pp. ???–???.

Bardet, G. (1920) Sur un syndrome d'obesite infantile avec polydactylie et retinite pigmentaire (contribution a l'etude des formes cliniques de l'obesite hypophysaire). PhD Thesis, name of institution, Paris, France.