

Research Article

## Assigning Level in Data-mining Exercises

Paul Hooley, Ian J. Chilton, Daron A. Fincham, Alan T. Burns and Michael P. Whitehead

School of Applied Sciences, University of Wolverhampton

Date received: 14/02/07

Date accepted: 16/03/07

---

### Abstract

*There is currently much interest in ascribing outcomes to Masters (M) level programmes. It is particularly difficult to define M level outcomes in bioinformatics for students on non-specialist programmes. An approach is described that attempts to discriminate undergraduate from M level in a data-mining exercise. Differentiation of level is based upon the taxonomic origin of a DNA sequence, the relative increase in gene complexity from lower to higher eukaryote and the initiative required to use a wider range of databases and analytical tools.*

**Keywords:** Masters, descriptors of level, data-mining, bioinformatics

---

### Introduction

Vast quantities of original research data, often generated by the genome projects, are available to the public via a variety of websites and can provide intriguing opportunities for novel teaching activities (Campbell, 2003). These now include user-friendly packages to teach bioinformatics, notably the European Multimedia Bioinformatics Educational Resource, EMBER (Attwood *et al.*, 2005). Data-mining uses computers to extract and analyse information to generate useful biological hypotheses and insights. We have previously described an exercise in data-mining that formed part of a final year undergraduate module. This used fungal DNA sequence data as the starting point for final year undergraduate (level 3) students on a range of B.Sc (Hons.) programmes to explore some of the tools available for bioinformatics (Hooley *et al.* 2004 available on British Mycological Society sponsored website; Fungi 4 Schools, <http://www.fungi4schools.org/>). Fungi have relatively short genes, for example *Aspergillus nidulans* has a mean gene length of less than 2000 base pairs, and many genes have only 1 or two introns (<http://www.broad.mit.edu/annotation/fungi/aspergillus/genome>). Nonetheless fungal genes illustrate all the major features of eukaryotic gene structure. In contrast human genes have a mean of 9 or 10 introns per gene making the typical human gene very much longer and commonly allowing for complex alternative splice reactions giving more than one protein product (IHGSC, 2004).

Information concerning an individual gene is referenced as an "accession number" on a publically available database such as NCBI (National Centre for Biotechnology Information, <http://www.ncbi.nlm.nih.gov/>). Fungal accessions were chosen to incorporate information relating to the entire gene structure,

generally including information on translation start and stop sites and intron number and position, as well as links to complete annotations on the relevant genome projects. Each student was randomly assigned a fungal accession number from NCBI. The exercise was carried out over an entire term, comprising 50% of the assessment on the final year undergraduate module Gene Manipulation and was based largely upon a workshop approach. Each student worked at their own pace on an individual accession number and asked for staff assistance as required. The final assessment was based upon a short report where the student accurately identified the gene, described the major features of its structure, and compared its similarity at both the DNA and encoded protein levels to its nearest relatives (Hooley *et al.*, 2004).

Having used this exercise successfully at undergraduate level we then considered how such an approach could be adopted for an audience of M level students from a wide variety of geographical and academic backgrounds (Table 1). The objective was to develop an M level data-mining module suitable for students on non-specialist and specialist bioinformatics courses.

**Table 1** Countries of origin of students registered on the MSc. DNA Data-mining module

Country of origin	2005/6	2006/7
UK	3	2
India	16	12
Nigeria	2	2
Ghana	1	2
Cameroon	0	1

#### **A Masters (M Level 4) Data-mining module**

It is difficult to find a common consensus for criteria that mark the move from B.Sc. (Hons.) to M level particularly in such a new area as bioinformatics. The U.K.'s Quality Assurance Agency for higher education have attempted to provide statements of outcomes as "threshold" or "good" for undergraduate work. They have also developed benchmarks for three subject areas at M level (business/management, engineering and pharmacy) and continue to work towards drafting a generic statement of M level outcomes (QAA, 2006, Bellingham pers. comm.). Hack and Kendall (2005) recently proposed some useful learning outcomes to distinguish between undergraduate, postgraduate (Masters in Bioscience and Masters in Bioinformatics) and PhD levels. These focused largely upon the move from undergraduate to M level involving the ability to produce biological models and a deeper understanding and application of analytical models and their parameters. The EMBER package used on the M.Sc. in Bioinformatics at the University of Manchester defines several basic tutorials with a smaller number of advanced tutorials and case studies.

One might expect M level material to be more challenging and open ended than undergraduate level, perhaps with a stronger element of original thought or analysis and incorporating the ability to negotiate the form of the particular assessment under study. We could not assume that our M level students had a common grounding in molecular biology (Table 1) so the teaching had to

consider an introductory approach which nonetheless led to work acceptable by its conclusion at an M level. Only a small number of our students were enrolled on a course in Molecular Biology and Bioinformatics which contains a significant component of computing modules. The majority specialised in Applied Microbiology and Biotechnology which has no compulsory computing components. Hence, as pointed out by Hack and Kendall (2005), whilst we can assume Masters students in Bioinformatics will be able to write programmes and develop databases, these skills are not expected in Masters students studying other Bioscience subjects. Although Honts (2003) reports that the successful introduction of simple programming skills at undergraduate level in Cell Biology courses at Drake University in the US was a useful addition to teaching bioinformatics to non-specialist students.

The M level module 'DNA Data-mining' begins with three formal lectures that cover revision of basic features of transcription and translation, the key tools of similarity comparison focusing upon the Basic Local Alignment Search Tool (BLAST, Altschul *et al.* 1990), the main genome databases and an introduction to protein structure modelling. This is followed by several tutorial exercises that utilise the application of BLAST in a simple worked example from a fungal gene to illustrate the key principles. Each student then works on an analysis of their own accession both in their own time and in a weekly timetabled slot in the computer lab with a member of staff present. The overall aim is for the student to provide as full an analysis as possible of an individual DNA sequence and the protein it encodes. In contrast to the undergraduate exercise, these sequences are selected from a collection of higher eukaryotic accessions on NCBI which do not contain complete details of gene structure and function. Students are supplied with a workbook and a support package on the university intranet, the Wolverhampton on line framework (Wolf) that logically takes them through the analytical tools that they may expect to use with a suggested time planner. For example this includes the application of easily used resources such as PSIPRED (McGuffin *et al.*, 2000) that allows the prediction of protein secondary structure and the prediction of potential transmembrane domains and more advanced programmes like Genthreader (Jones, 1999) which attempts protein fold predictions. Multiple sequence alignment programs for comparison of several DNA or protein sequences together e.g. ClustalX (available from European Bioinformatics Institute, EBI, <http://www.ebi.ac.uk/>) and 3D-Coffee (Wallace *et al.*, 2005) may be employed. They were guided to specific tools freely available on some sites such as the EMBER tutorials (Attwood *et al.*, 2005) initially available on line, now as a CD also, or less structured but more comprehensive resources like EBI. Some reports then conclude with the design of phylogenetic trees to explore the evolutionary relationships of their accession (Baldauf, 2003, Hall, 2004). Weekly surgery tutorial workshops are provided where staff are on hand to deal with individual problems and suggest appropriate analytical approaches. The final module assessment is based around the submission of a 5,000 word report with Figures, Tables, Legends, etc. being additional to this limit to encourage a concise reporting style.

## Defining M Level Features

Table 2 compares details available from two separate example accession numbers – one representing a relatively simple fungal gene and the other a human DNA sequence. The entire gene structure can be modelled from a single paper and the accession itself for the fungal gene. In contrast the human gene gives multiple references and the sequence is mRNA derived and therefore lacks genomic annotation.

**Table 2** A comparison of data available from a selected fungal and a human accession number at NCBI

Feature	AF202995	NM_152854
Origin	Complete genomic DNA clone	cDNA derived from mRNA
Species	<i>Aspergillus nidulans</i>	<i>Homo sapiens</i>
Gene Product	Zn finger transcription factor	CD40/ Tumour necrosis factor receptor
Papers referenced	1	10
DNA sequence length b.p.	3814	1554
Introns	2 with exact positions given	No detailed information but two alternative splice products mentioned
Upstream activating sequence (promoter)	Complete	Incomplete

The student will have to access other databases (such as ensembl: <http://www.ensembl.org/>) to acquire gene structure details. This may involve renewed Blast searches on different sites as NCBI accession numbers may not be recognised on a foreign database. For example searches of the OMIM (Online Mendelian Inheritance in Man) database will be required. One could also choose higher eukaryotic accessions where no clear functions had been ascribed and no papers were referenced. The following features then mark a move from an undergraduate exercise to an M level.

- Use of accession numbers that relate only to non-genomic gene sequences – i.e. cDNA's generated from reverse transcription of mRNA molecules. Some initiative was therefore needed in employing a wider range of search tools and databases to collate analyses of complete gene sequences and their encoded products.
- Use of higher eukaryotic genes i.e. from animals and plants. These may have a very complex structure showing greater size and a large number of introns with alternative splicing reactions also possible to generate more than one potential protein product. Students were given a wide range of choice of accession number, often reflecting their own wider interests or research project work.
- Performing phylogenetic analysis on the DNA or protein obtained would be viewed as indicative of excellence at undergraduate level. However at M level this was seen as a standard form of analysis to undertake.
- Students are encouraged to use more than one analytical approach to scrutinise any inconsistencies in predictions, for example of secondary

structures of proteins encoded by their genes. This fits Hack and Kendall's (2005) suggestion of the need for a good understanding of the underlying parameters of models and their statistical bases.

- The mark scheme rewards curiosity – where students have followed an unusual or ingenious method of analysis, perhaps using less conventional databases or altering search tools away from their default settings. Criticisms of the shortcomings of the tools are encouraged. Whilst the tools and websites introduced in the early weeks are relatively straightforward to use, other analytical tools require a more sophisticated level of understanding by the student. This supports another of Hack and Kendall's (2005) criteria for M level implying the use of a wider range of analytical tools.

Presently we have only chosen conventional protein-encoding genes for analysis so avoiding genes encoding rRNA, tRNA or micro RNA (Thadani and Tammi 2006) as the final product. However as our knowledge of the RNA world increases, teaching a data-mining approach to the analysis of such genes could follow a similar pattern.

**Table 3** Web sites and programs used by a student who undertook a data-mining task at both undergraduate (fungal gene) and M (human gene) levels

Program/database	Description	URL (Uniform Resource Locator)
<i>Undergraduate assignment</i>		
Rps-Blast (NCBI)	Conserved domain search	<a href="http://www.ncbi.nlm.nih.gov/BLAST">www.ncbi.nlm.nih.gov/BLAST</a>
Blastn (NCBI)	DNA sequence homology search	<a href="http://www.ncbi.nlm.nih.gov/BLAST">www.ncbi.nlm.nih.gov/BLAST</a>
Blastx (NCBI)	Translated DNA sequence homology search	<a href="http://www.ncbi.nlm.nih.gov/BLAST">www.ncbi.nlm.nih.gov/BLAST</a>
<i>Postgraduate assignment</i>		
Rps-Blast (NCBI)	Conserved domain search	<a href="http://www.ncbi.nlm.nih.gov/BLAST">www.ncbi.nlm.nih.gov/BLAST</a>
Blastn (NCBI)	DNA sequence homology search	<a href="http://www.ncbi.nlm.nih.gov/BLAST">www.ncbi.nlm.nih.gov/BLAST</a>
Blastx (NCBI)	Translated DNA sequence homology search	<a href="http://www.ncbi.nlm.nih.gov/BLAST">www.ncbi.nlm.nih.gov/BLAST</a>
InterPro database (EBI)	Taxonomic coverage analysis	<a href="http://www.ebi.ac.uk/interpro">www.ebi.ac.uk/interpro</a>
ClustalW (EBI)	Phylogeny analysis	<a href="http://www.ebi.ac.uk">www.ebi.ac.uk</a>
Ensembl database	Primary structure	<a href="http://www.ensembl.org">www.ensembl.org</a>
PSIPRED	Secondary structure prediction	<a href="http://bioinf.cs.ucl.ac.uk/psipred/psiform.html">bioinf.cs.ucl.ac.uk/psipred/psiform.html</a>
MEMSAT 2	Transmembrane topology prediction	<a href="http://bioinf.cs.ucl.ac.uk/psipred/psiform.html">bioinf.cs.ucl.ac.uk/psipred/psiform.html</a>
GenTHREADER	Tertiary structure prediction	<a href="http://bioinf.cs.ucl.ac.uk/psipred/psiform.html">bioinf.cs.ucl.ac.uk/psipred/psiform.html</a>
UniProt (Universal Protein Database)	Domains	<a href="http://www.pir.uniprot.org/">www.pir.uniprot.org/</a>
CATH Protein Structure Classification Database	Domain architecture	<a href="http://cathwww.biochem.ucl.ac.uk/">cathwww.biochem.ucl.ac.uk/</a>

An example of how M level can be observed is by comparing the work of one of the authors (I.J.C.) who undertook these bioinformatics assignments both as an undergraduate and postgraduate student. For the undergraduate exercise (on the honours year module ‘Gene Manipulation’), the 1000 word limit resulted in an 11-page report. The M level (DNA Data-mining module) 5000 word limit resulted in a 34 page report; 12 and 24 references were cited respectively. Table 3 outlines the databases and programs that the student used. This highlights the far greater degree of independent investigation undertaken by the student at M level.

## Evaluation

The student-centred format of the module quickly allows the student to find their existing level, identify their shortcomings and seek help on a one to one basis with staff. Conversely as suggested by Campbell (2003), gifted students can rise to the challenge to provide work of exceptional quality. The chances of plagiarism are reduced by each student being independently assigned a separate sequence for analysis with weekly sessions with staff providing an “informal viva” environment so that personal student input and overall understanding are monitored. Whilst some supporting literature references are expected in the report, the bulk of the text represents the student’s own unique analysis so reducing the “cut and paste” culture. This workshop approach is therefore quite demanding of staff time when compared to a conventional lecture format.

**Table 4** A summary of module evaluation forms from three separate iterations of the M level DNA Data-mining module (24 forms returned from 58 distributed)

Statement	Responses			
	<i>Excessive</i>	<i>About Right</i>	<i>Light</i>	<i>No Response</i>
The amount of directed reading	<b>10</b>	<b>12</b>	<b>1</b>	<b>1</b>
The volume of assessment	<b>6</b>	<b>18</b>	<b>0</b>	<b>0</b>
	<i>High</i>	<i>Medium</i>	<i>Low</i>	
The degree of difficulty compared to other modules	<b>10</b>	<b>14</b>	<b>0</b>	
	<i>Yes</i>	<i>No</i>	<i>No Response</i>	
Did you find tutorials/workshops helpful?	<b>19</b>	<b>3</b>	<b>2</b>	
Did you find the learning experience stimulating?	<b>18</b>	<b>6</b>	<b>0</b>	
Would you recommend the module to other students?	<b>19</b>	<b>3</b>	<b>2</b>	

M level modules at the University of Wolverhampton are scored on a non linear grade scheme, A – D for pass marks, E for a marginal fail and F for a fail. Of 58 students taking the module over 3 separate iterations in 3 years, the mean grade has been C for each year with a mean pass mark of 83%, 84% and 95%. The numbers of students enrolled on the M.Sc. in Molecular Biology and Bioinformatics were too small to compare statistically with the major cohort studying the M.Sc. in Applied Microbiology and Biotechnology but it was noted that these students, as one might expect, tended to score higher than the average (for example in one year : A, A and B, compared with a module mean of C). Table 4 summarises the responses of students to a voluntary questionnaire at the end of the module. Interestingly they were aware of the heavy load of outside reading required with approaching half the respondents considering this excessive. Nonetheless the actual assessment loading was considered reasonable by a large majority of students. Just under half the students found the degree of difficulty high compared to other M level modules (most of which comprised a traditional diet of teacher led lectures and tutorials). Most encouragingly a large majority of students found the learning experience stimulating and would recommend the module to their peers.

**Table 5** Performance on M level modules accessed by the 2005/2006 cohort to compare numbers of students and grades on DNA Data-mining with four other modules assessed by conventional exams and continuous coursework. Total module numbers vary depending upon some flexibility in individual student programmes with the Gene Technology module including a large number of additional students studying the M.Sc. in Biomedical Sciences

Title	Module Grade					
	A	B	C	D	E	F
Genes & Genomes	0	4	6	8	3	6
DNA Data-mining	3	6	4	7	0	1
Masters Lab.Techniques	3	7	9	3	2	1
Research Methods	0	12	8	4	2	4
Gene Technology	6	17	22	5	6	2

Table 5 shows the relative performance of students on the DNA Data-mining module compared with: Masters Laboratory Techniques assessed by continuous assessment of course work; Research Methods, which uses a multi component, task-based assessment regime; Genes and Genomes culminating in a substantial piece of coursework handed in at the end of the course; and Gene Technology, which employs a traditional three hour examination. The performance of students on DNA Data-mining compares favourably with those modules that use a variety of conventional assessment tools.

## Conclusions

There is disagreement over the ability to provide useful benchmark statements of a generic nature at M level or for subject specific outcomes at this level. For example some awards may be based around professional bodies' requirements and M.Phil. programmes will have a more distinctive research flavour than taught programmes (QAA 2006). Nonetheless we

suggest that within individual exercises or modules M level achievements can be discriminated. We propose that a useful criterion to apply to teaching data-mining at undergraduate versus M level relies upon the taxonomic origin and hence the complexity of the genes used in exercises, particularly for students that are not on specialist bioinformatics programmes. Fungal genes can provide ideal and simple models of eukaryotic gene structure where remarkably complete analyses can be accomplished with access to a minimum number of websites. Genes from higher eukaryotes can often provide more challenging projects better suited to M level work that require a deeper understanding of gene and protein structure as well as the concepts of phylogenetic relationships. They also allow the student to express the ability to think creatively in the use of a diverse range of websites and databases.

### Acknowledgements

We would like to thank Dr Eldridge Buultjens of Abertay University for his many helpful suggestions and support in the establishment of our teaching programmes in bioinformatics. We thank the University of Wolverhampton Celt initiative for support.

### Communicating Author

Dr Paul Hooley, School of Applied Sciences, University of Wolverhampton, Wolverhampton, West Midlands. WV1 1SB. Tel: 01902 322667 Fax: 01902 322714  
Email: [p.hooley@wlv.ac.uk](mailto:p.hooley@wlv.ac.uk)

### References

- Altschul, S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410
- Attwood T.K., Selimas I., Buis R., Altenburg R., Herzog R., Ledent V., Ghita V., Fernandes P., Marques I. and Brugman M. (2005) Report on the EMBER project – a European multimedia bioinformatics educational resource. *Bioscience Education e-Journal* **6**  
<http://www.bioscience.heacademy.ac.uk/journal/vol6/Beej-6-4.htm> (last accessed 16/03/2007)
- Baldauf S.L. (2003) Phylogeny for the faint of heart ; a tutorial. *Trends in Genetics* **19**, 345–351
- Campbell A.M. (2003) Public access for teaching genomics, proteomics and bioinformatics. *Cell Biology Education* **2**, 98–111
- Hack C. and Kendall G. 2005 Bioinformatics education in the UK : are we educating scientists or training technicians? *Centre for Bioscience Bulletin* **15** , 10–11
- Hall B.G. (2004) *Phylogenetic Trees Made Easy – A How-To Manual*. Sinauer Associates Inc.; US
- Hooley P., Burns A.T.H. and Whitehead M.P. (2004) Fungal gene sequences make excellent models for teaching data-mining. *The Mycologist* **18**, 118–124

- Honts J.E. (2003) Evolving strategies for the incorporation of bioinformatics within the undergraduate cell biology curriculum. *Cell Biology Education* **2**, 233–247
- International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945
- Jones D.T. (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology* **287**, 797–815
- McGuffin L.J., Bryson K. and Jones D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405
- QAA (2006) *Securing and maintaining academic standards : benchmarking M level programmes.*  
<http://www.qaa.ac.uk/academicinfrastructure/benchmark/masters/MlevelbenchmarkingFeb06.pdf> (accessed 16/03/2007)
- Thadani R. and Tammi M.T. (2006) *BMC Bioinformatics* **7** (Suppl. 5 ):S20
- Wallace I.M., Blackshields G. and Higgins D.G. (2005) Multiple sequence alignments. *Current Opinion in Structural Biology* **15**, 261–266