*Mycologist, Volume **18**, Part 3 August 2004. ©Cambridge University Press Printed in the United Kingdom.*
DOI: 10.1017/S0269915XO4003040

Fungal gene sequences make excellent models for teaching data mining

PAUL HOOLEY, ALAN BURNS & MICHAEL WHITEHEAD

School of Applied Sciences, University of Wolverhampton, Wulfruna St., Wolverhampton, West Midlands. WV1 1SB. UK. Email P.Hooley@wlv.ac.uk, Fax +44 01902 322714

A brief introductory exercise in the use of on-line databases to examine fungal genes and their products is described. Fungal genes make particularly good teaching models owing to their relatively simple eukaryotic structure and wide range of homologues in higher organisms including humans. An evaluation of students' reactions to the exercise is included.

Keywords: Data mining, BLAST, Fungal genes

Background

Due to the almost exponential increase in DNA sequence data there is now a pressing need to teach undergraduate students how to manipulate and analyse its information content (Dyer & LeBlanc, 2002). This process of "data mining" makes use of computers to extract and analyse information concealed within large data sets. The provision of large quantities of original research data in the public domain by the genome projects offers exciting new opportunities for teaching activities (Campbell, 2003).

Gene manipulation (module code AB3024) is a level 3 (final year undergraduate) module taught as part of a modular degree scheme at the University of Wolverhampton. Recently we have incorporated an exercise in data mining to introduce students to the use of on-line databases. All modules are constructed around a set of outcomes which are skills that the students are expected to be able to demonstrate at the end of the activity. One of the specific outcomes for this module is: "by successful completion of the module, students will be able to appreciate that advances in bioinformatic areas can provide full molecular details of genes and gene products". In addition there are transferable skills such as the use of information technology associated with this module. This exercise will then contribute to the achievement of these outcomes. The module is populated by a diverse range of students on biological sciences based degree programmes. These encompass mature students as well as a proportion of overseas students for whom English is a second language. Students may have only a modest information technology background. Hence this teaching activity is aimed at biology lecturers employing pre-existing software tools so reducing the need to utilise colleagues with a formal computing background.

Fungal genes are particularly useful for teaching undergraduate DNA data manipulation techniques as the genes tend to be small (1-5kb) and therefore manageable, yet contain a number of eukaryotic gene structures (Table 1) which will enable the student to explore and understand genes from a range of organisms. Other organisms (like humans) can provide genes which are of unmanageable size and complexity for many undergraduates.

Table 1 Examples of features that may be identified in fungal genes

TATA boxes
CAAT boxes
Protein binding regions
Start codons
Signal sequences
Cleavage sites
Introns (5', lariats, 3')
Exons
Stop codons
Polyadenylation signals

With careful selection of the fungal genes a variety of concepts can be introduced to the students. These concepts include the fact that perhaps 50% of human proteins have a fungal homologue making fungal models especially valuable for understanding their

human counterparts and disease processes (Zeng *et al*., 2001). DNA sequences are stored on databases as "accession numbers". Ascomycete isopenicillin N synthase (e.g. accession number X17436) can prompt an investigation of hypotheses of horizontal transfer from streptomycete bacteria based on DNA sequence similarity (Walton, 2000). Fungal polygalacturonases (e.g. X64356) can illustrate gene families, orthologues (genes in two species derived from a single gene in the last common ancestor), paralogues (genes formed by duplication of a sequence in one species) and the similarity of encoded proteins with plant products involved in fruit maturation (Bussink *et al*., 1992; Torki *et al*., 2000) . Polyketide synthases (e.g. AY495605) provide examples of genes with complex evolutionary origins (Kroken *et al*. 2003). Concepts of lineage specific genes (or orphans) such as genes controlling sporulation can be considered by comparison to eukaryotic wide products like cell cycle control components. Thus this whole assignment has the advantage of introducing students to a variety of fungal species and products that they may not have previously encountered.

Table 2 gives examples of some useful web sites relevant for the analysis of fungal DNA sequences. There are a number of tools available for rapid comparisons of similarity of sequences of DNA or protein. Perhaps the most widely used tool is BLAST (Basic Local Alignment Search Tool, Altschul *et al*., 1990). A range of useful general texts which introduce this tool includes Brown (2000) and Lesk (2002) whilst a guide to internet-based tools is provided by Fortna & Gardiner (2001). The activity takes place over the bulk of a university semester, around nine weeks from initiation to the final submission of a report. It is embedded within a formal lecture series covering techniques of gene cloning and characterisation.

The session begins with an empirical one hour revision tutorial where students are asked to manually read a Sanger's dideoxy DNA sequencing gel and translate this in one of the possible reading frames into a potential gene product using the standard genetic code table. Again manually they then provide an estimate of the similarity between their DNA and protein sequences and a second DNA and protein sequence provided for them. This encourages them to understand the concepts of reading frames and codon redundancy. It also gives the students an appreciation of the practical problems of interpreting sequencing results.

The second tutorial involves the students moving to the computer lab to begin the task shown in the worked example section which follows below. This task employs BLAST to carry out the same activity that they have laboriously performed manually but using a much larger database.

The worked example centres around the NCBI (National Centre for Biotechnology Information) site but similar exercises can be attempted with a range of databases (Table 2). This example focuses on the analysis of a DNA sequence cloned from *Aspergillus nidulans* (O'Neil *et al.*, 2002) - a gene chosen to inform the students of research being performed in their own department.

Worked example

It is now straightforward to perform a variety of sophisticated internet based searches to compare DNA and/or protein sequences and so suggest likely functions. The most comprehensive database is at NCBI (www.ncbi.nlm.nih.gov/). Typically the accession number for a DNA or protein molecule may also include a relevant journal reference, locus details and information relating to gene structure such as intron positions.

How do we identify the likely function encoded by a DNA sequence? The simplest approach might be to look for similarities in the given sequence with previously characterised molecules. Given that the genetic code can also predict potential gene products we can also compare similarities in proposed proteins encoded by the DNA too.

- You can open the accession at the NCBI site by simply typing **AF202995** in the box at the top of the home page (on the pull down menu - Search nucleotide for:)
- Once you have obtained your sequence you can then

Table 2 Examples of URLs of interest for fungal DNA sequence analysis

Database	URL
National Centre for Biotechnology Information	http://www.ncbi.nlm.nih.gov/
Sanger Centre	http://www.sanger.ac.uk/
Neurospora crassa Shear & Dodge genome Project	http://biology.unm.edu/biology/ngp/home.html
Saccharomyces cerevisiae Hansen genome project	http://www.yeastgenome.org/
Aspergillus nidulans (Edam) Winter genome project	http://www.broad.mit.edu/annotation/fungi/aspergillus/
Candida albicans (Robin) Berkhout genome project	http://genolist.pasteur.fr/CandidaDB/
Fungal Genome Resource	http://gene.genetics.uga.edu/

```
1     actgcagtga  gacagcttgg  ccagcagaaa  gcctctcaca  gcagggctgt  ggccgatgac
61    atggcattct  gacatgagag  aacggtcagg  ctagctgagg  tccaagcatt  actgagagcg
121   ccattgagag  ggcggatcat  agtgcgggat  tgttcaggca  acacaaatct  gcgatattcc
181   attatcaaga  aacgagccca  tttgatatgc  gatggccagc  attgtgagcg  gcgcgtatgc
241   attctcgtcc  aatagaaaca  ctgccatcgt  tatcgaacgt  ttgtggctgt  gacgacaagg
301   cgatgctgtg  tctgacgtcg  gcttcaagac  ctctctgtcc  cgtcctctcc  tcacctcatg
361   tatttattat  gctcgcagtt  cagcatcttc  aaattctgag  agtaatccca  gtaagcatac
421   tttgatgcaa  tcgactgcgg  tcttgagggt  ttataatgga  tcgatgatcg  gtgcgtctcg
481   aacaccgcca  tgatggtgga  gttgcagttg  ggccgatgat  ctgccatccg  gagaaaagcc
541   gtaaattagg  aaattgattg  ctattttcgt  acctttttct  ttaccacggt  gtcttcaaag
601   cttgcagtca  ttggttatac  ctgagacctg  cggggctatg  aatacgagga  aaagcgtgcc
661   gaccacggat  gaacacgggt  tccgtgtatg  ggcgtcgtaa  taatggccta  aatccgaagt
721   aaaatacaaa  aaacactgcc  ccagagaccc  gtcgtgtatc  cggtaatacg  gccccttggt
781   ctttcctttt  agggattagc  acatctcgtc  tatcaggcag  acgtccaaga  cgcctcagcc
841   tcaatccatc  atcaagaagg  catcacccgc  tcttcccgac  cctgatctta  tcccatccca
901   ttccgtctct  tttcccaagt  ctccatcgtt  cacgtccaag  agtcgtgttt  tgtttggacc
961   ctctaggttc  cagcattttt  ttatattcat  acacatcgaa  cccaactttc  caccctcccc
1021  ttttcacctt  tcccaccgtc  agtctgttgt  tctcccgtca  gcgctcaggg  cgtcacctgc
1081  tgttgacaga  gagtccgacc  tctgtatgta  tttcatcaac  gccccgtcgc  ccgaccgtct
1141  acgaagagtc  gatatattcc  caaccttgct  cttgtgtaca  ttctaacaat  cgtcaagtcg
1201  gcagacctgc  caatcgttga  ctggctgcac  cacttcccca  tctaatcaga  agcttccgtt
1261  ggatttggtc  acgatcagtg  agtcatttac  ctgcatcact  ctgccggcct  gtccttgcta
1321  acacctttтc  ggacattagg  gaccgtccat  c**atg**agtcca  gcacaagact  ctgagtctat
1381  aaaggcccac  ccgaggcgcc  ggcccttcag  ggcagcccgg  ccctcgcttg  tcccagctga
1441  agagcaatca  ccgtcactac  ctccacttcg  tctacgcaaa  ggtgaaacat  ttaatccatc
1501  tattcttcgc  tcttccgacc  gtgaccacct  tgtaccgtcg  ctgccacgcc  gatctcccac
1561  atgccctggc  gctctggaag  ctatcgccgc  tggacaacag  cgtatggccg  acatcttgga
1621  gcgccttgac  ttgaactcgg  gtaccacgtc  cacctccgat  gaaaacgacg  acctccccgt
1681  ccctaagggt  ttacttcggc  ttcatttaca  aactcaagca  cggagagagg  gcaccgttga
1741  accgcattcg  cgccaaccta  gcccgatgcc  caaggaacat  tctcggaagg  ctcagagagt
1801  ccattgtcat  gcttctgata  gtggaattgg  ctcctctatc  agcagtgctc  aatccgtgtc
1861  gtctaacaaa  ggtacatggt  catggacagc  tctttggtaa  gcatgagcta  acaatttcac
1921  agtgaaagcg  ggacaattgt  ctcgtagcaa  cctcccaacc  tcccgttccc  agtccgccat
1981  tacccgctcc  atcagtgcca  tggatgccca  aagcactcag  cgacacaaac  tcagctctga
2041  gggtcgagcc  gagatcgaaa  agcacgtcat  cggccctctt  ctagaggatg  agaaatcgaa
2101  gccgttccac  cctattcttg  aagatgtccg  tcagcaaatt  gacgatgaac  gtatctcctg
2161  cctgcgtgac  cttgaaaaaa  cagtattctc  gctcgctccg  gaggtgaaga  caaatgatgc
2221  tgcctatgtt  cgattctgcc  agtataccat  cctgtgcctc  ggccagacgg  tctcattcct
2281  caatggccgg  gacctgtgcc  tgccgactga  taagcagtac  aacaacggtt  acttcgtgga
2341  cctcttagac  caggtttctc  aattcaagag  aatccgtgac  gaatggaaga  gaaggcacga
2401  ggctgacggc  aaggtcaagt  atgtaatctc  cttttgcatc  gtattccaac  agttactaac
2461  cgtctaccat  agggctccgc  aacttagact  cgagggtgga  ctctcccaaa  ccggacgcct
2521  tcttgaaatg  gtcgttgagc  aggacggcga  ggctatttcc  ctgcgtacag  gcaaacctta
2581  tgaaggtcag  cctattcctt  cgatgaagcg  atctctcagc  gcggcttcta  ccgacgaagg
2641  agttcagcgt  tccatggccc  gtcgcaagaa  gaatgcgcct  cctatgaaca  tcaacaagaa
2701  gtgcaaggac  tgtgacaagg  tctttgctcg  gccatgcgac  ctgacaaagc  acgaaaaatc
2761  tcacagtcgt  cccttcaaat  gccctgttac  ttcctgcaag  taccatatca  agggctgggc
2821  cactgagaaa  gaatcagagc  gtcactacaa  tgacaagcac  tctgatgcac  cgcgcctctt
2881  tgcttgtcaa  tttgagtcct  gctcctacaa  gtctaagaga  gaaagcaact  gcaagcaaca
2941  catggagaag  actcatggtt  gggtgtacat  gcgatctaag  aacaacggaa  gaagtaaggc
3001  gtctcctcag  caacagacta  cctcgccgtc  cagcagctcg  gttcagccca  agcaggctcc
3061  ctccgtctgg  agcatgacac  ctccttccga  agcgcctgac  taccggcagg  agccgaatgg
3121  ctgggacctt  gccccctctc  cggagactcc  ggatttgttc  aacacctatc  aagctcccat
3181  gactgccatg  cctggctcag  tgaccggaac  attggatgcc  gtcaccccaa  ctacagggac
3241  catcaactct  ccgtctgaac  cgttcgatct  tgcgcaagag  aatacagcct  tttctattca
3301  ggatatattc  ccggaaatga  aggctagtga  cggtttgttg  tttcctggcg  gagacatgga
3361  ttaccctgat  ttcatcaaca  accacaacat  gttcaatgac  tttggctacg  gtgatttcac
3421  catgccgaca  caagggttgc  aatatggaga  gacacagcaa  cctcaatttg  aagatgactc
3481  ggctggcttc  ctcctcgatg  tctacaacga  catgcacacg  tacgggatta  accccggccc
3541  tggtggtctc  **taa**tctcgtc  acatttggtg  gttgagacag  accccggtgg  ttgaccatct
3601  tgatcaatgt  gttacgaccc  tacctatttg  caggcaaatg  tttattggct  cctttcgcga
3661  aaataagttgg  ctacattgcg  ctgtaatata  ctgcaccact  ctggataggg  cagataagat
3721  aattcatcgt  tttcgctatg  ttttcttctg  ttaaaattaa  gcataagaaa  agtcgctcct
3781  acggttgagc  ttcgtagaga  accttatcct  gcag
```

Fig 1 An example of the annotation shown to students of the DNA sequence of accession number AF202995. Start and stop codons are shown in bold with intervening exons underlined. Potential poly A signal sites are shaded. Data on intron positions may be confirmed experimentally by cDNA analysis and detailed in the accession itself. Internet based programmes such as the open reading frame finder at NCBI can be used by the students themselves to check the accuracy of such annotations. Similarly students may use lecture notes to identify potential polyadenylation sites or consensus sequences within promoters. In producing such figures students should use a font which gives a consistent character size such as Courier.

perform a homology search using BLAST (Basic Local/Linear Alignment Search Tool) click on:

- BLAST (top of NCBI home page)
- Standard nucleotide – nucleotide BLAST (BLASTN).
- You may then type or cut and paste the sequence into the Search box. Alternatively at NCBI merely typing the accession number will suffice – this easy option does not apply to most databases because of the inability of one database to 'talk' to another and recognise its shorthand. Leave other settings on the default values (conventional levels of sensitivity for a specified number of matches).
- Click BLAST (at the bottom left hand side of page) to start the search.
- After a few moments you will be given an I.D. number to check the file for your completed search. Such files are generally kept on the server for up to a week. Remember that at busy times (normally in the afternoon in the UK) the file will take some time to process so note the file number and do something else!

Whilst waiting for the files students are encouraged to work through the excellent on-line tutorials on the use and interpretation of BLAST.

The third tutorial takes place the following week when all students are expected to have created a file for analysis of the worked example. They should have read the information provided on the accession number page and performed BLASTN and BLASTX (DNA translated in each of the six possible reading frames matched against a protein database) searches for the worked example. The student should prepare questions for the tutor concerning any details they do not understand.

Example results

The results file consists of a summary Fig demonstrating the distribution of matching 'hits' to the query sequence, a Table summarising the accession numbers of such hits, and individual comparisons of bases or amino acids for the best matches.

A direct examination of the DNA sequence of the accession along with its 'annotation' (labelling of relevant features) may allow some structural features of the gene to be identified (Table 1). A typical example of a fungal gene sequence which provides evidence of some of these structures is shown in Figure 1.

This example of a BLAST search will demonstrate a limited number of specific filamentous fungal homologues and also the widespread occurrence of a DNA binding domain in eukaryotes. Students were encouraged to create files for both DNA/DNA searches and translated DNA or protein/protein searches. The completed files should include example summary tables of the closest matches on the database. Table 3 shows a summary of the best matches for a DNA v DNA search for the worked example. Note that the first match is effectively a control with the sequence compared to itself. In this case only the second match provides a confident 'hit' and a possible identification of part of the query sequence.

A search at the encoded protein level for most accessions will generally give many more significant matches because of the redundancy/degeneracy built into the DNA genetic code where more than one codon can encode the same amino acid (Table 4). Figure 2 shows an example from a BLAST analysis of a detailed protein match using single letter code. Students can explore the detail of this further, by colour coding groups of amino acids into for example small non-polar (GAST), hydrophobic (CVILPFYMW), polar (NQH), negatively charged (D,E) or positively charged (K,R) residues. Once a consistent match is achieved for a particular gene product the students can begin a conventional literature search or online database searches using for example Pubmed (found at NCBI

Table 3 A summary of DNA sequence similarities to AF202995. Score (bits) gives a measure of the number of matches in the raw alignment and E values provide statistical measures of confidence. For example an E value of 1 means that one match like this would be expected by chance in a database of this size. The larger the bit score and the smaller the E value the more confident we are of the match.

Sequences producing significant alignments:	Score (bits)	E Value
gi|14195702|gb|AF202995.2| *Aspergillus nidulans* strain R153...	7519	0.0
gi|22726229|gb|AY072919.2| *Talaromyces emersonii* zinc finge...	52	0.007
gi|27731040|ref|XM_218473.1| *Rattus norvegicus* similar to z...	44	1.7
gi|54000|emb|X00229.1|MMRNO3 Mouse tRNA gene cluster for tR...	44	1.7
gi|54904|emb|X07460.1|MMTRND15 Mouse mAsp1 DNA flanking tRN...	44	1.7
gi|6604548|gb|AC006968.2|AC006968 *Homo sapiens* PAC clone RP...	44	1.7
gi|21322181|gb|AC002078.2| *Homo sapiens* BAC clone CTB-111H1...	42	6.6
gi|19033958|gb|AC007239.3| *Homo sapiens* BAC clone RP11-83A1...	42	6.6
gi|13897301|emb|AL390334.4|CNS06C7N Human chromosome 14 DNA...	42	6.6
gi|2809270|gb|AC002349.1|AC002349 *Homo sapiens* Xp22 PAC RPC...	42	6.6

Table 4 A summary of translated DNA to protein database similarities for AF202995 (BLASTX, 149 hits in total)

Sequences producing significant alignments:	Score (bits)	E Value
gi|14195703|gb|AAF15889.2| putative zinc finger transcripti...	1288	0.0
gi|33115142|gb|AAL69549.3| zinc finger transcription factor...	464	e-129
gi|32699313|sp|Q9P8W3|ACE1_TRIRE Zinc finger transcription ..	302	e-80
gi|32423175|ref|XP_332025.1| hypothetical protein [*Neurospo*...	285	5e-75
gi|38104048|gb|EAA50669.1| hypothetical protein MG04428.4 [..	77	1e-72
gi|32411121|ref|XP_326041.1| hypothetical protein [*Neurospo*...	156	e-36
gi|38110713|gb|EAA56393.1| hypothetical protein MG06364.4 [...	122	5e-26
gi|40744315|gb|EAA63491.1| predicted protein [*Aspergillus n*...	102	6e-20
gi|40745993|gb|EAA65149.1| hypothetical protein AN0644.2 [A...	58	1e-06
gi|19577366|emb|CAD28447.1| putative zinc finger transcript...	58	2e-06

site). Protein searches at NCBI are also automatically processed through the Conserved Domain Database (CDD) which may place the query protein into a recognised family. In this case the gene product is clearly identified as a transcription factor belonging to the common C_2H_2 zinc finger class.

Exercise and assessment

The students are provided with the following assessment criteria to perform the assignment.

- **Produce a report that describes the likely function of this gene and its homologues.**
- A specific accession number will be assigned to each individual student.
- You are provided with the accession number for a DNA sequence from the NCBI database. From the information provided on the database page give structural information about the gene. Using whatever databases you think relevant compare this gene and its product to other characterised genes and their products. From this describe the role of this gene, the family of proteins to which its product belongs and any homologous genes from other species. **1000 words + Tables, Figures, references.**

Grade A: A comprehensive and concise report which includes details covering gene structure such as introns/promoters/terminators, relationship to other genes and comprehensive analysis of the gene product using on-line facilities.

Grade D (pass): A legible report that correctly describes gene function and protein and DNA homology to other sequences.

In the following weeks nine further hours of computer lab time are built into the module timetable. These are not formally taught by staff but are rather workshops where each student can discuss progress on their own analysis with an individual member of staff.

Post-module evaluation

The relatively brief format (1000 words of text) with an onus on the production of Figures and Tables reduced the workload on staff with each report taking around 15 minutes to mark. The mark scheme was relatively simple rewarding correct identification of the gene, the ability to use and present BLAST results, incorporation of conventional reference sources and 'curiosity' including inquisitive explorations of the search tools and settings. Undergraduate modules at Wolverhampton University, like many other

```
Query: 2702   CKDCDKVFARPCDLTKHEKSHSRPFKCPVTSCKYHIKGWATEKESERHYNDKHSDAPRLF 2881
              C DC KVFARPCDL KH KSH+RPFKC + CKY  GW T KE ERH NDKH+  P ++
Sbjct: 20    CPDCTKVFARPCDLNKHSKSHTRPFKCLHSDCKYADLGWPTLKELERHNNDKHAPNPIIY 79

Query: 2882   ACQFESCSYKSKRESNCKQHMEKTHGWVYMRSKNNGR 2992
              AC++E C YKSKRESNCKQHMEK HGW+Y RSK+NG+
Sbjct: 80    ACEYEGCDYKSKRESNCKQHMEKAHGWLYTRSKSNGK 116
```

Fig 2 A comparison of the protein encoded by AF202995 with a match from *Neurospora crassa*. The + symbol denotes an amino acid of similar size/shape/charge whilst identical matches are shown directly. This area of similarity represents a DNA binding domain common to many eukaryotes characterised by pairs of cysteines and histidines (in bold) that chelate zinc ions to form a finger- like structure. Three zinc fingers are implicated here.

universities, are marked using a non-linear 16-grade system known as the Common Grade Point System. A percentage mark of 50-52 would correspond to a C8 grade, 53-56 to a C9 grade, whilst A16 covers the 80-100% mark range.

Student evaluation of modules is obtained by providing them with questionnaires that allow student anonymity, containing 21 specific machine-readable questions where the student fills in an appropriate box. In addition the reverse side of the questionnaire allows students room to write individual comments e.g. suggestions on how the module could be run another time, the best/least satisfactory things about the module etc. Twenty-eight students took the opportunity to fill in the non-compulsory questionnaire. Given the individual accession numbers and subsequently highly individual Gene Reports, that reflect both the differing bioinformatic details available for a particular gene as well as the students' ability, it is not surprising, in this new venture, that 8 of the students (approx. 28%) mentioned in some form or other that they would have liked further guidelines and explanations on the Gene Report. The responses to the most relevant specific questions, set out in Table 5, reveal that the workload was not excessive, that the degree of difficulty was not excessive, and the module overall was highly popular. In addition the mean student score (n = 31) for the Gene Report was C9.38 whilst that for the written exam component was slightly lower at C9.16. A group of international students (5) to whom bioinformatics was totally novel also scored better in the Gene Report (C = 9.2) compared to their exam performance of C8.2. In summary there is no evidence to support the concept that this class of students were disadvantaged by the introduction of this bioinformatics assignment. Students in general appeared to have coped well with the Gene Report.

Individual assignments with each student investigating a different fungal gene certainly reduced the possibility of plagiarism between students. It was a useful way to discover if a student had misunderstandings in basic molecular biology, for example in appreciating the implications of a DNA sequence encoding six possible reading frames. The BLAST results may show both paralogues and orthologues although it was usually only the best students who would pick up on these concepts. Some new concepts could be introduced, such as moonlighting proteins – where a match to part of a protein suggests a function for this domain only with perhaps additional functions residing elsewhere (Jeffery, 1999). This concept is particularly well illustrated by the BLAST approach where the extent of similarity across a comparison of two or more genes or proteins is obvious. General shortcomings of databases such as their mutual incompatibility and the use of different confusing formats should be impressed upon students. Similarly it was possible to begin to educate students on the likely error sources in the databases themselves and the analytical tools they employ. These might include simple DNA sequencing errors or incorrect annotations as well as more intriguing phenomena such as where alternative splicing of transcripts may give rise to more than one protein from a given gene (Peri *et al*. 2001).

Perhaps there is an attraction in this type of activity for some foreign students where a small quantity of original writing in English was demanded and marks were awarded for analytical skills. The activity emphasised the phenomenon of data overload – students had little problem in creating files but found it difficult to abstract the relevant information. There was the temptation to deluge the students with useful websites and on the second iteration of this activity the number of web sites initially recommended was greatly reduced from around fifteen to just two. It was often difficult to encourage all the students to attend the drop-in workshops and students frequently wanted to see staff outside their timetabled hours. The exercise represented a marked change in teaching style for the students with the onus being upon them to show initiative and this allowed independent students to excel. Campbell (2003) gives examples of how good

Table 5 Responses from a selected range of questions asked on a questionnaire (containing 21 queries) returned from (28) students on the evaluation of module AB3024, University of Wolverhampton

	Excessive	**About right**	**Light**
Was the volume of assessment during the module?	28.57%	71.43%	0%
	High	**Medium**	**Low**
What was the degree of difficulty, compared with other modules?	7.14%	89.29%	3.57%
	No	**Yes**	
Would you recommend the module to other students?	14.29%	85.71%	

students can adopt quite different pathways to the analysis of *Saccharomyces cerevisiae* DNA sequences by using BLAST searches as a prompt for examining conventional literature sources.

Conclusions

There is a temptation to under-resource activities such as this, naively believing that all that is required is a collection of working computers. This can underestimate the importance of maintaining a decent staff/student ratio (1 : 10 minimum) and effectively limits the class size (McInerney, 2003). Staff were often surprised by examples of computer-phobic students (where even 'cutting and pasting' proved a challenge) and the reluctance of students to use web-based tutorials. Most students required regular face to face contact with staff if only for reassurance. Nonetheless there is much scope for talented students to be challenged. The background knowledge of and interest in mycology improved amongst many students reflecting a belief that a broader subject understanding can be gained by data mining approaches (Campbell, 2003).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403 – 410.

Brown, S.M. (2000). *Bioinformatics : a biologist's guide to biocomputing and the internet*. Eaton Publishing.

Bussink, H.J.D., Buxton, F.P., Fraaye, B.A., Degraaff, L.H. & Visser, J. (1992). The polygalacturonases of *Aspergillus niger* are encoded by a family of diverged genes. *European Journal of Biochemistry*. **208**: 83-90.

Campbell, A.M. (2003). Public access for teaching genomics, proteomics and bioinformatics. *Cell Biology Education* **2**: 98 – 111.

Dyer, B.D. & LeBlanc, M.D. (2002). Meeting Report : Incorporating genomics research into undergraduate curricula. *Cell Biology Education* **1**: 101 – 104.

Fortna, A. & Gardiner, K. (2001). Genomic sequence analysis tools : a user's guide. *Trends in Genetics* **17**: 158 - 164.

Jeffery, C.J. (1999). Moonlighting proteins. *Trends in Biochemical Sciences* **24**: 8 - 11.

Kroken, S., Glass, N.L., Taylor, J.W., Yoder, O.C., & Turgeon, B.G. (2003). Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proceedings of the National Academy of Sciences*. **100**: 15670-15675.

Lesk, A.M. (2002). *Introduction to Bioinformatics*. Oxford University Press.

McInerney, J.O. (2003). A bioinformatics teaching programme : do as I say and not as I've done. *Society for General Microbiology 153rd Meeting UMIST Proceedings Abstracts* 19 - 20.

O'Neil, J., Bugno, M., Stanley, M.S., Barham-Morris, J.B., Woodcock, N.A., Clement, D.J., Clipson, N.J.W., Whitehead, M.P., Fincham, D.A. & Hooley, P. (2002). Cloning of a novel gene encoding a C_2H_2 zinc finger protein that alleviates sensitivity to abiotic stresses in *Aspergillus nidulans*. *Mycological Research* **106**: 491 498.

Peri, S., Ibarrola, N., Blagoev, B., Mann, M. & Pandey, A. (2001). Common pitfalls in bioinformatics - based analyses : look before you leap. *Trends in Genetics* **17**: 541 - 545.

Torki, M., Mandaron, P., Mache, R. and Falconet, D. (2000). Characterization of a ubiquitous expressed gene family encoding polygalacturonase in *Arabidopsis thaliana*. *Gene*. **242**: 427-436.

Walton, J.D. (2000). Horizontal gene transfer and the evolution of secondary metabolite gene clusters in fungi : an hypothesis. *Fungal Genetics and Biology* **30**: 167 - 170.

Zeng, Q., Morales, A.J. & Cottarel, G. (2001). Fungi and humans : closer than you think. *Trends in Genetics* **17**: 682 - 684.