

Making bioinformatics projects a meaningful experience in an undergraduate biotechnology or biomedical science programme

Iain C. Sutcliffe and Stephen P. Cummings

*Biomolecular and Biomedical Research Centre, School of Applied Science,
Northumbria University*

Date received: 19/06/2007

Date accepted: 19/07/2007

Abstract

Bioinformatics has emerged as an important discipline within the biological sciences that allows scientists to decipher and manage the vast quantities of data (such as genome sequences) that are now available. Consequently, there is an obvious need to provide graduates in biosciences with generic, transferable skills in bioinformatics. We present here an example of how bioinformatics work can be developed, using bioinformatics tools freely available over the internet, to provide a challenging and worthwhile honours project experience for undergraduates undertaking degree programmes in biotechnology and biomedical sciences. We argue that this type of project work can provide an appropriate, stimulating alternative to bench-based ('wet') laboratory projects. Such projects develop skills that complement and extend traditional laboratory skills.

Keywords: *Bioinformatics; Genomics; Lipoproteins; Phylogenetics*

Introduction

Since the publication of the first microbial genome sequence in 1995 (Fleischmann *et al.* 1995), the past decade has seen an explosion in biological sequence data. For example, as of July 2007, the GOLD database (<http://www.genomesonline.org/gold.cgi>; Liolios *et al.*, 2006) lists 626 completed published genome projects and 2129 ongoing genome projects (excluding metagenomes!). This vast volume of data (and that derived from the post-genomic technologies that have subsequently emerged, such as microarray technologies) represents both a wonderful resource but also a considerable and rapidly growing challenge to the biological scientist. As noted in a recent editorial (Delpech, 2006) this information is not *per se* knowledge and so the discipline of bioinformatics has emerged to aid in the collation, presentation and deciphering of this information. Moreover, many primary data publications contain some degree of bioinformatic content that must be critically evaluated (for example, sequence alignments and phylogenetic trees). It is thus increasingly important that pedagogy in the biological sciences keeps abreast of developments in this area and that students are educated in the handling and critical evaluation of such data. It is relevant to distinguish between the needs to educate specialist bioinformaticians for the development of bioinformatics tools and the need to educate the end-users of those tools (Gollery, 2006). A recent report identified this problem with the delivery of bioinformatics teaching on many bioscience courses. In most cases students are taught how to utilise tools without reference to the development of these resources. This is unsurprising as the latter requires extensive multi-disciplinary training in both computer science and biology (Likić, 2005), an approach which is typically outwith the scope of a generalist bioscience undergraduate curriculum. While it is a laudable aim to address this problem, most bioscientists remain users of bioinformatics rather than practitioners and will, as

a result, increasingly require the skills to use the rapidly increasing amount of bioinformatics data effectively.

However, as the amount of available software increases, and the underlying algorithms become more complex, there is a danger of training 'bioinformatics technicians': graduates with the ability to use a number of 'standard' bioinformatics tools; rather than life scientists with the capacity to identify the appropriate tool to solve particular problems and understand their advantages and limitations (Hack and Kendall, 2005). With this in mind, we present here an example of how bioinformatics-based work can be incorporated into final year undergraduate biosciences honours projects.

Previous research has demonstrated that providing opportunities for students to undertake individual projects promotes the acquisition of deep, context specific knowledge, challenges them to solve problems and make decisions (Millenbah and Millspaugh, 2003) and promotes critical reflection and discussion of the research activity they are engaged with (Seifert, 2004; McCune and Hounsell, 2005). The bioinformatic projects we offer have microbiological themes and relate to study programmes in biotechnology and biomedical sciences. Students in biotechnology have received prior education in bioinformatics through 10 credit specialist modules in their first- and second years of study whereas for students in biomedical sciences these projects are their first significant experience of bioinformatics. We contend that such projects represent a valuable alternative or complement to 'wet' laboratory projects, with excellent opportunities for students to demonstrate their grasp of theories and concepts and apply them to the problem they are studying (Millenbah and Millspaugh, 2003). With careful design, such projects offer sufficient scope and flexibility to readily enable the differentiation of the most academically able and engaged students. Moreover, it allows the students to utilise transferable skills that need to be part of the graduates toolkit, such as the ability to critically analyse and process complex data sets and contextualise their work in the biology of the organism(s) they are studying (Seifert, 2004). This example will hopefully illustrate the relevance and potential of this type of project work for educators in both the higher and further education sectors.

Project example: Bioinformatic analyses of lipoproteins encoded in bacterial genomes.

Bacterial lipoproteins are a family of proteins that can be readily identified by the bioinformatic analysis of conserved sequence motifs (Sutcliffe and Harrington, 2002; Babu et al, 2006). Project work in this area was designed based on the authors experience of bacterial genome sequence analysis (Sutcliffe and Harrington, 2002; Sutcliffe and Harrington 2004a, 2004b; Sutcliffe and Hutchings, 2007). The projects are delivered in the form of weekly individual or group tutorials, supported by emailing of instructions, protocols and targets to the students. The students then have to complete their bioinformatic analyses as independent study, reporting back in subsequent tutorials and eventually completing a project report (typically ca. 5000 words) for assessment. Specifically, a starting dataset for a given published bacterial genome can be rapidly retrieved by use of a pattern search strategy (Gattiker et al, 2002) and then the sequences features of the individual protein sequences in the dataset re-evaluated using a variety of open access internet tools for signal peptide analysis, bacterial lipoprotein identification and predictors of membrane protein topology (Figure 1). Data is collated in excel files and thus the students need to demonstrate effective management of electronic stored data. Overall, retention of a sequence in the dataset is determined using a 'majority vote' approach (Tjalsma and van Dijk, 2005). Thus students are able to exclude sequences considered to be 'false-positives' (typically ca. 10% of the total) from their datasets. Subsequently, potential 'false negatives' can be recovered using alternative pattern searches and, if necessary, alternative strategies such as keyword searches of the genome annotation and reference to the published

literature. These potential ‘false negatives’ are also evaluated to exclude false-positives. This data reduction process provides a powerful introduction to the need to critically evaluate the outputs from bioinformatics tools and especially to the need to evaluate different bioinformatic tools in comparison to one another. Todd and Weaver (2005), highlighted the value of students exploring the outcomes of analysing datasets using a variety of tools in reinforcing that these tools are essentially predictive relying on different algorithms with inherent assumptions that require further analyses.

Putative Lipoproteins in a genome identified by ScanProsite pattern search analysis with a conserved amino acid pattern (‘lipobox’ features)

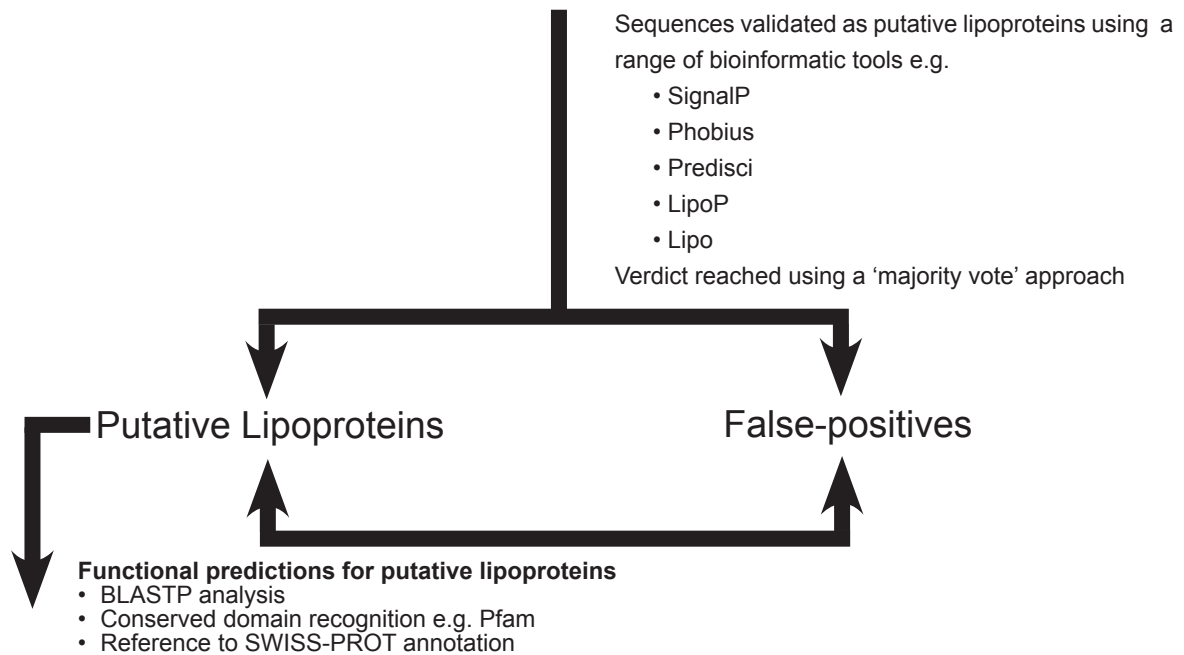


Figure 1 Flow diagram illustrating the structure of a typical bioinformatics project. Note that, since putative lipoproteins typically represent ca. 2% of a given proteome, projects can be scaled to match different sized assessments by choosing genomes of different sizes or by allocation of only a proportion of the sequences recovered by pattern searching.

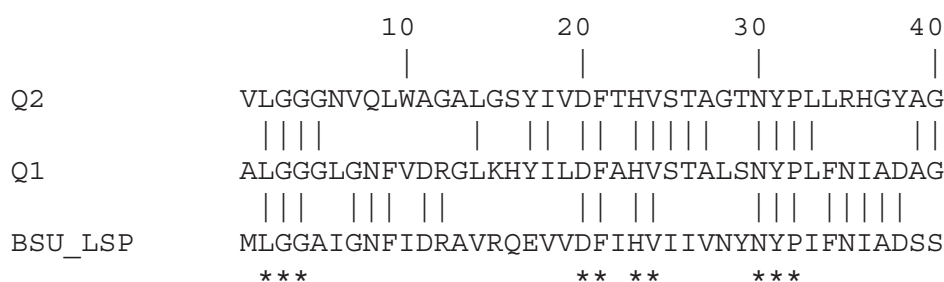


Figure 2 Sequence alignment illustrating the problem of ‘putativism’. Sequence BSU_LSP is a 40 amino acid stretch of the Bacillus subtilis lipoprotein signal peptidase enzyme, covering the important conserved aspartate catalytic diad (Tjalsma et al., 1999). Theoretical sequence Q1 is 20/40 (50%) amino acids identical to BSU_LSP and might reasonably be annotated as a putative lipoprotein signal peptidase, since the crucial aspartates (▲) are conserved. Theoretical sequence Q2 is highly homologous to Q1 (20/40, 50% amino acid identity) and all three sequences are homologous to each other (* below alignment indicate amino acid identity). However, even though sequence Q2 is homologous to BSU_LSP (12/40, 30% amino acid identity) the critical aspartate catalytic diad is not conserved and it would be erroneous to annotate this sequence as a putative lipoprotein signal peptidase. This illustrates the flaw in the logic that because A is similar to B and B is similar to C, then A must have the same function as C: proteins A and C may have diverged to the point where they are functionally distinct. Allowing students to explore these concepts can be a valuable learning experience.

Table 1 Analysis of bioinformatic project marks compared to overall cohort mark over three academic years

Academic Year	Total students in cohort	Project mark $\bar{x} \pm SD$	Bioinformatics projects	Project mark $\bar{x} \pm SD$
2004/5	38	59 \pm 12	3	60 \pm 5.5
2005/6	40	57 \pm 9	3	59 \pm 14
2006/7	42	60 \pm 8	4	61 \pm 6.4
Totals	120	59 \pm 10	10	60 \pm 8.6

The second stage in the projects allows the students to perform functional analyses on individual proteins retained in their datasets. The annotation made available in UniProtKB/SWISSPROT (us.expasy.org/sprot/; ref) and the links therein to outputs from other databases (e.g. Pfam) can be collated with the outputs from BLAST homology searches (carried out using the NCBI server www.ncbi.nlm.nih.gov/BLAST/). Interpreting the outputs of these searches is extremely challenging and develops the students confidence and initiative. In particular, the need to distinguish between significant homologues that are hypothetical proteins versus significant homologues that are experimentally characterised proteins can be reinforced. This also provides a powerful introduction to the concept of 'putativism' (Wassenaar and Gastra, 2001) and the need to critically evaluate sequence homology data (Figure 2).

This can lead to an important introduction to concepts such as orthology and paralogy, as well as provide examples of probable gene duplication events. Finally, students may identify through context specific understanding and their literature searches that lipoproteins are frequently involved in ABC transporters, where they occur with a membrane permease and an ATP binding protein, or in two component systems, as an accessory protein along with a membrane bound sensor kinase and a DNA binding response regulator,. Therefore they can test the robustness of the lipoprotein identification by studying the chromosomal context of the protein of interest.

These studies often alert students to problems of mis-annotation (putativism), which can result from uncritical copying over of information during automated computer based annotation. Indeed, the authors own experience in this kind of analysis has already lead to the re-annotation of several sequences in the SWISS-PROT database, thereby benefiting the general scientific community. Bioinformatic studies such as these also encourage independent exploration: most online bioinformatics tools and outputs are replete with hyperlinks to other databases, servers and information sources. Collating and deciphering this information presents an excellent and sometimes daunting challenge, through which the student can be guided in tutorials. In our project module students are awarded 20% of the marks by the supervisor by demonstrating initiative and independence, their ability to plan and record their work and the implementation of their work. The tutorials are initially a mechanism to explain the process and allow the student a forum to ask questions and seek clarification on the process. However, as the project develops they become increasingly student led enabling the supervisor to differentiate students through their engagement with these criteria. Furthermore, the students are challenged to interpret their findings in the specific context of the biology and scientific interest (e.g. biotechnological use or pathogenicity) of the organism being studied. Because the members of this protein family have a broad range of functions, the students are also introduced to a wide range of biological concepts and systems (for example, bacterial membrane transport systems, cell surface enzymes, cytochrome biology, protein export systems, the significance of the taxonomic distribution of conserved hypothetical proteins e.t.c.). Finally, these functional analyses can be

used to teach students how their bioinformatics work can be used to formulate new hypotheses and influence experimental design. Indeed, it is important to emphasise throughout that bioinformatics is fundamentally a predictive discipline and a foundation but not a substitute for experimental enquiry. These features thus give the students considerable opportunity to demonstrate the depth of their knowledge and understanding. Consequently, following the completion of the project reports, the profiles for this type of project vary across the full marks range but correlate well with the student's ability, engagement and scientific sophistication and are comparable with those of the remainder of the cohort undertaking conventional laboratory based projects (Table 1). Student feedback suggests that the students enjoyed these project and found them as challenging as the work they observed their peers doing in wet labs. Indeed, several students have been sufficiently motivated by the projects to consider continuing with MSc studies in bioinformatics.

Conclusion

The specific example above illustrates both the level of challenge and the rewards that can be embedded in bioinformatics-based project work. More generally, the analysis of proteins families represented in genomes allows the supervisor considerable versatility since these projects can be adapted by varying the type of protein family to be analysed (and the number of proteins contained therein), either through selection of genomes of appropriate size or the allocations of only a proportion of the proteome to particular students. Most notably, these projects provide undergraduates with transferable skills in many aspects of data handling and the confidence to proceed with bioinformatics work. These skills are of considerable utility to the modern biologist, given that many experimental hypotheses are framed from bioinformatics analyses. Nevertheless, the merits of providing 'dry' bioinformatics work rather than 'wet' laboratory work may be considered contentious. However, there have been concerns raised over the value of project work due to its expense in both staff time and the resources required to support this activity (McCune and Hounsell, 2005). Obviously, projects that are solely computer-based preclude undergraduates from gaining knowledge of techniques that could be experienced in laboratory based projects. On the other hand they do deliver in a comparable way in terms of developing deep knowledge, critical reflection, the ability to apply concepts and solve problems in the same way as laboratory based studies. They also prevent less able or engaged students from "throwing expensive chemicals down the sink" (Wood, 2004). It is also often the case that the laboratory skills and techniques taught in projects can be highly specialised, whilst more generic transferable skills are typically very capably reinforced in the taught modules of biotechnology and biomedical sciences programmes. Qualitative feedback indicates that students completing bioinformatics projects find them to be enjoyable rewarding and challenging educational experiences. To address the issues of the types of the skill sets that 'wet' and 'dry' projects tend to emphasise, we are now developing projects that extensively mix bioinformatic analyses with practical work, such as the design of degenerate primers for PCR and their application in the amplification of novel 16S rRNA sequences for phylogenetic analyses. Clearly bioinformatic analyses can be built into a variety of project experiences to embrace the changing practices of biologists following the molecular biological and genomics revolutions. This aim is consistent with the recent comment of Gollery "What is most needed are biologists who have a basic understanding of bioinformatics. A combination of wet lab skills and computational expertise will make a powerful combination" (Gollery, 2006).

Communicating Author

Dr Iain C. Sutcliffe, Biomolecular and Biomedical Research Centre, School of Applied Science, Northumbria University, Newcastle upon Tyne NE1 8ST, U.K. Phone: 0191 227 3176 Email: iain.sutcliffe@unn.ac.uk

References

- Babu, M. M., Priya, M. L., Selvan, A. T., Madera, M., Gough, J., Aravind, L. and Sankaran, K. (2006) A database of bacterial lipoproteins (DOLOP) with functional assignments to predicted lipoproteins. *Journal of Bacteriology* **188**, 2761-2773
- Delpech, R. (2006) Bioinformatics and school biology. *Journal of Biological Education* **40**, 147-148
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., Mckenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. and Venter, J. C. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496-512
- Gattiker, A., Gasteiger, E. and Bairoch, A. (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. *Applied Bioinformatics* **1**, 107-108
- Gollery, M. (2006) An assessment of the current state of bioinformatics education. *Bioinformatics* **1**, 247
- Likić, V. A. (2006) Computer programming and biomolecular structure studies. *Biochemistry and Molecular Biology Education* **34**, 1-4
- Liolios, K., Tavernarakis, N., Hugenholtz, P. and Kyripides, N. C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Research* **34** (Database Issue), D332-D334
- McCune, V. and Hounsell, D. (2005) The development of students' ways of thinking and practising in three final-year biology courses. *Higher Education* **49** (3), 255-289
- Millenbah, K. F. and Millspaugh J. J. (2003) Using experiential learning in wildlife courses to improve retention, problem solving, and decision-making. *Wildlife Society B* **31**, 127-137
- Seifert, T. L. (2004) Understanding student motivation. *Educational Research* **46**, 137-149
- Sutcliffe, I. C. and Harrington, D. J. (2002) Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. *Microbiology* **148**, 2065-2077
- Sutcliffe, I. C. and Harrington, D. J. (2004) Lipoproteins of *Mycobacterium tuberculosis*: an abundant and functionally diverse class of cell envelope components. *FEMS Microbiology Reviews* **28**, 645-659
- Sutcliffe, I. C. and Harrington, D. J. (2004). Putative lipoproteins of *Streptococcus agalactiae* identified by bioinformatic genome analysis. *Antonie van Leeuwenhoek* **85**, 305-315
- Sutcliffe, I.C. and M.I. Hutchings (2007). Putative lipoproteins identified by bioinformatic genome analysis of *Leifsonia xyli subsp. xyli*, the causative agent of sugarcane ratoon stunting disease. *Molecular Plant Pathology* **8**, 121-128
- Tjalsma, H. and van Dijk, J. M. (2005) Proteomics-based consensus prediction of protein retention in a bacterial membrane. *Proteomics* **5**, 4472-4482
- Tjalsma, H., Zanen, G., Venema, G., Bron, S. and van Dijk, J. M. (1999) The potential active site of the lipoprotein-specific (type II) signal peptidase of *Bacillus subtilis*. *Journal of Biological Chemistry* **274**, 28191-28197
- Wassenaar, T. M. and Gastra, W. (2001) Bacterial virulence: can we draw the line? *FEMS Microbiology Letters* **201**, 1-7
- Weaver, T. and Cooper, S. (2005) Exploring protein function and evolution using free online bioinformatics tools. *Biochemistry and Molecular Biology Education* **33**, 319-322
- Wood, E. (2004) "Why offer final year projects?" Higher Education Academy Centre for Bioscience Conference - Making the Most of Final Year Projects. <ftp://www.bioscience.heacademy.ac.uk/events/cardiff/wood.pdf> (date visited 21/ 05/ 07)