There are many different tools of the bioinformatics trade. The technical languages in which computer programs are written are often used as tools to take the monotony out of routine tasks; database management tools are used to help organise data more efficiently; statistical tools are used to provide robust data analyses; and so on.

This topic guide will look at a range of computational tools used for analysing biological data, with a particular focus on those used to analyse protein sequences.

On successful completion of this topic you will:
• be able to carry out data analysis within the field of bioinformatics (LO4).

To achieve a Pass in this unit you need to show that you can:
• apply mathematical and statistical methods to analyse biological data (4.1)
• use software tools for efficient pattern detection (4.2)
• present information clearly (4.3).
1 Introduction to analysis tools

Many different data analysis tools are used within the field of bioinformatics. For example:

- Perl® and Python™ are popular high-level scripting languages used in bioinformatics; Java™ and C++ are object-oriented languages
- Microsoft Excel® is a spreadsheet application, with calculation and graphing tools and tables, often used to create simple databases; MySQL™ and PostGreSQL are relational- and object-relational database management systems that offer more sophisticated tools for data storage and maintenance
- R and Bioconductor are tools that offer comprehensive statistical and programming environments, especially for life-science data analyses (including microarray, sequence and genome analysis).

No matter what the task, it is important to choose the right tool for the job. Simple solutions (flat-files, Excel®, Perl®) are often appealing, but might not always give ideal results. Often, it is worth investing extra time to understand and reap the benefits of using more elaborate tools.

2 Computational tools for analysing biological data

The methods outlined previously underpin some of the key tools that are used to analyse biological data. We saw how sliding-window algorithms can reveal the most hydrophobic and hydrophilic regions of sequences, and how dotplots can provide effective ‘visual’ overviews of pairwise sequence alignments. We also saw that robust pairwise-alignment algorithms have had to be optimised for use as swift database search tools: the most popular programs in this latter category are FastA and BLAST.

Key terms

- Sliding-window algorithm: An algorithm that performs calculations in a window of pre-defined length that runs incrementally over a set of values (such as successive amino acid hydrophobicity values in a protein sequence), often used to plot a series of averaged scores, such as in graphs of amino acid properties.
- Alignment tool: A program that takes biological sequences as input, and outputs them so that residues that are identical or similar are in vertical register, or alignment.

Activity

Plot a hydrophobicity profile for human rhodopsin:

1. enter the identifier, OPSD_HUMAN, into the ProtScale web-form [http://web.expasy.org/protscale/](http://web.expasy.org/protscale/) in the box that asks for a UniProtKB/Swiss-Prot® ID
2. scroll down to the bottom of the page and change the window size to 17, then click on Submit
3. on the next page, click on Submit again.

- How many transmembrane domains are there in the protein? Verify your answer by examining the Sequence Annotation Features in the protein’s UniProtKB/Swiss-Prot® entry [http://www.uniprot.org/uniprot/P08100](http://www.uniprot.org/uniprot/P08100).

We saw that dynamic programming algorithms are computationally prohibitive for aligning multiple sequences, so heuristics are used to approximate the best alignment. Most alignment tools thus use a heuristic algorithm that progressively aligns sequences in pairs, following the branching order of a family tree: example implementations of this approach include ClustalW, T-Coffee and MUSCLE.
Different alignment programs give different results, so tools have been developed to allow automatically-generated alignments to be refined manually: examples of such alignment editors are CINEMA and Jalview. Editors are valuable, not just for allowing alignment errors to be corrected, but also for providing a range of additional functionalities: for example, allowing sequence characteristics documented in database Feature Tables to be viewed in an alignment context; or features of alignments to be related to structural characteristics (if 3D coordinates are available).

Take it further
Find out more about producing and analysing sequence alignments in Understanding Bioinformatics (Zvelebil and Baum, 2007), Garland Science.

Choosing the right tool for the job

It is important to understand what analysis tools were designed to do, and how their input parameters affect their results. One vital ‘input parameter’ is the sequence itself. Earlier, we examined the bloated flat-file format used to store sequences in databases; more streamlined sequence formats are necessary for efficient computation. The most common is FastA, a crude derivative of the NBRF-PIR format. When any analysis tool requests an input sequence, it is important to choose the right format, or the tool will either fail or give unpredictable results.

Activity
Examine the FastA format (http://en.wikipedia.org/wiki/FASTA_format) and consider also the NBRF-PIR format (http://www.bioinformatics.nl/tools/crab_pir.html).
• What are the main differences between them?

There are many other input parameters, depending on the nature of the analysis task. For instance, when using a database search tool (for example, BLAST), a database and scoring matrix relevant to the required search must be selected – the default parameters will always give a result, but this may not be the most appropriate result for the search you intended.

Similarly, when using alignment tools, the scoring matrices they use, the insertion/deletion penalties they apply, and the evolutionary assumptions they make must be considered – comparing results from different tools is then more meaningful. It is also important to understand the colour scheme used to display the results. Colour schemes have a crucial influence on how alignments are interpreted – the best tools allow the default scheme to be changed to allow better understanding of what alignments mean.

Likewise, when searching protein family databases, the level of specificity of the diagnosis must be considered: whether it is to detect members of a superfamily or of a specific subfamily, or simply to identify a functional site, it is important to select the database with the right level of granularity.
Activity
Reflect for a moment on what we learned about the methods that underpin the protein family databases, PROSITE, PRINTS and Pfam.
• Which database would be most suitable for diagnosing the presence of specific functional sites in a query sequence?
• Which would be most suitable for diagnosing its subfamily membership?
• Explain your reasoning.

4 Seeking consensus between different tools

As we have seen, many analysis tools do similar things, and it is important to understand how they and their results differ. It is not that one particular tool is best – no tool is infallible; rather, it is a question of using the right tool or combination of tools for the task at hand. In almost every area of bioinformatics, experience shows that combining multiple approaches produces more reliable results than do individual approaches themselves.

Consider the large number of alignment tools. To remove the need for users to have to choose which program to use, M-Coffee combines ~8 methods into a single package. It computes the alignments from all these tools, combining them into a consensus result. The consensus is believed to be ‘better’ than the result of the best individual method, most of the time.

Similarly, think of the proliferation of protein family databases. To try to simplify matters, InterPro integrates ~12 such databases and provides unified views of family membership and domain organisation for any query sequence. The database is unique in allowing sequences to be analysed simultaneously from multiple perspectives, giving consensus results that are more informative than those from its member databases alone.

Activity
Examine the InterPro entry for ephrin A5: http://www.ebi.ac.uk/interpro/protein/P52803.
• From the table of matched sequence features, which results from a match to a PRINTS fingerprint and which from a match to a PROSITE conserved site?
• To which fingerprint motifs does the PROSITE match correspond? (Note: be sure to set the entry type to ‘Family’ and colour by ‘source database’).
George works for a small software company. His role is to develop user-friendly tools for protein sequence analysis. His primary project involves the design of a software suite that integrates several different analysis tools within a single interface, to optimise the ease-of-use for customers. The software suite, known as Utopia (http://utopia.cs.man.ac.uk/utopia/), allows users to align protein sequences, to view their annotations from UniProt, to visualise their 3D structures, and so on, all within the same package: http://utopia.cs.man.ac.uk/utopia/2mhr-structure-and-annotations.

Software suites like this make the work of database curators, and of other users interested in protein sequence analysis, much easier, more efficient and more reliable. This is because such packages remove the need for users to have to switch between lots of different individual analysis tools, and to convert data formats between them; moreover, they allow users to more quickly and effectively spot where there is consensus between the constituent analysis methods.

Further reading
Find out more about current methods for building integrated analysis tools from the Utopia website, http://utopia.cs.man.ac.uk/utopia/about-0.

Checklist
At the end of this topic guide you should be familiar with the following ideas about bioinformatics:
✓ many tools are available for analysing biological data, including programming languages and database management systems, database interrogation and search programs, statistical tools, alignment and visualisation tools, etc.
✓ different tools implement different methods and algorithms, even those designed to perform similar tasks
✓ understanding which tool to choose for a particular task is vital because different tools produce different results – choosing the right tool allows the output to be interpreted with greater confidence
✓ there is no single ‘best tool’ for any given task – different analysis tools have different strengths and weaknesses that affect how well they perform
✓ best results are usually obtained by using several tools and taking the consensus result from their various outputs.

Acknowledgements
The publisher would like to thank the following for their kind permission to reproduce their photographs:
\textbf{PhotoDisc}: Lawrence Lawry
All other images © Pearson Education
Every effort has been made to trace the copyright holders and we apologise in advance for any unintentional omissions. We would be pleased to insert the appropriate acknowledgement in any subsequent edition of this publication.