Joan C. Bartlett
McGill University, Montreal

Tomasz Neugebauer
Concordia University, Montreal

Supporting Information Tasks with User-Centred System Design: The development of an interface supporting bioinformatics analysis

Abstract: We present an interface to support the integration of bioinformatics analysis with scientific practice. The interface guides scientists through the co-ordinated use of a wide range of analyses and resources in order to solve a complex information task.

Résumé: Nous présentons une interface qui utilise l'intégration de l'analyse bio-informatique dans le domaine scientifique. Cette interface guide les scientifiques à travers l'utilisation coordonnée d'une vaste gamme d'analyses et de ressources, de manière à effectuer une tâche informationnelle complexe.

1 Introduction

Among the challenges people face navigating today's vast array of information resources is identifying what type of information is available, where and how to obtain it, and what to do with the information once it has been obtained. This is particularly so for those trying to accomplish a complex, goal-directed information task.

This paper presents the final phase of a larger study addressing a specific information task, that of linking bioinformatics analysis to scientific practice. Based on an integrated information behaviour and task analysis approach, this work has already produced a systematic protocol detailing the application of bioinformatics analysis to the scientific problem of predicting gene function from genetic sequence data. Here, we present the development of an interface designed to make the bioinformatics analysis protocol accessible and usable to a laboratory scientist.

2 Background

Bioinformatics has been defined as "the computer-assisted data management discipline that helps us gather, analyse, and represent [biological] information" (Persidis, 1999, 828). It had its genesis with the development of automated DNA and protein sequencing techniques in the 1970s, and the creation of computer-based, remotely accessible, central repositories of sequence information in the 1980s (Persidis, 1999). Since then, there has been an exponential growth of bioinformatics resources, consisting of databases of biological information (e.g., *GenBank, SwissProt*), and software tools (e.g., *BLAST*, *ClustalW*) that access, manipulate and analyse the data. The 2006 annual database issue of the journal *Nucleic Acids Research* lists 858 individual tools (Galperin 2006).

In order to integrate bioinformatics analysis into a scientific research problem or agenda, a laboratory biologist typically needs to use more than one bioinformatics tool. Yet, information about how to link several different bioinformatics analyses into a cohesive, integrated approach to solving a particular type of problem is not readily available. Anecdotal evidence suggests that it is this process that is problematic for laboratory scientists (B. Muskat, personal communication, November, 2000). Biologists tend to use "only the simplest tools available" (Butler 2001). The challenge is to know what type of information bioinformatics analysis can provide, which resources provide what information, and how the resources are used. This knowledge tends to be passed on by word of mouth. While there is a call for "a better understanding from the general biologist of the real possibilities given by the analysis of genomes" (Andrade 2003, 217), and the need "to integrate very tightly the bioinformatics with doing experiments"(Hood, as cited in Butler 2001), the bioinformatics literature has taken a different approach. It has tended to focus on individual techniques (e.g., Baxevanis and Ouellette 2001), the use of a specific resource (e.g., Baxevanis and Davison 2002), or the development and refinement of bioinformatics tools (e.g., Dalkilic and Costello 2004).

2.1 Scientific Scenario

The information task at the heart of this research is the functional analysis of a gene sequence, that is, predicting the possible or likely function of a gene product, based on its sequence data. This is a very timely problem, given the fact that the Human Genome Project and other large sequencing projects have generated vast quantities of sequence data, for which little or nothing is known about the biological significance or function. Determining the function of these genes is one of the major challenges for biomedical research. From a practical standpoint, laboratory determination of gene function may take weeks, months or even years. In contrast, using bioinformatics analysis to predict the function can take as little as a few hours. While the bioinformatics analysis does not provide a conclusive answer – the findings must ultimately be empirically verified in the laboratory – it is extremely valuable in guiding and directing the laboratory investigation in the most promising direction, ultimately leading to savings of both time and resources. Thus, this was a complex problem which had multiple sub-tasks involving many forms of data and information.

2.2 Bioinformatics Analysis Protocol

The bioinformatics analysis protocol was developed by synthesizing and integrating the individual approaches taken by a cohort of twenty bioinformatics experts to the problem of conducting a functional analysis of a gene sequence (Bartlett 2004; 2005). The protocol was validated by the original participants, and a new cohort of eighteen additional experts. The protocol describes a series of twelve analytical steps, grouped into three alternate pathways (see Figure 1).

The protocol has as its starting point the genetic (DNA) sequence data. The first three steps are preparatory, getting the data into the correct format for further analysis. These

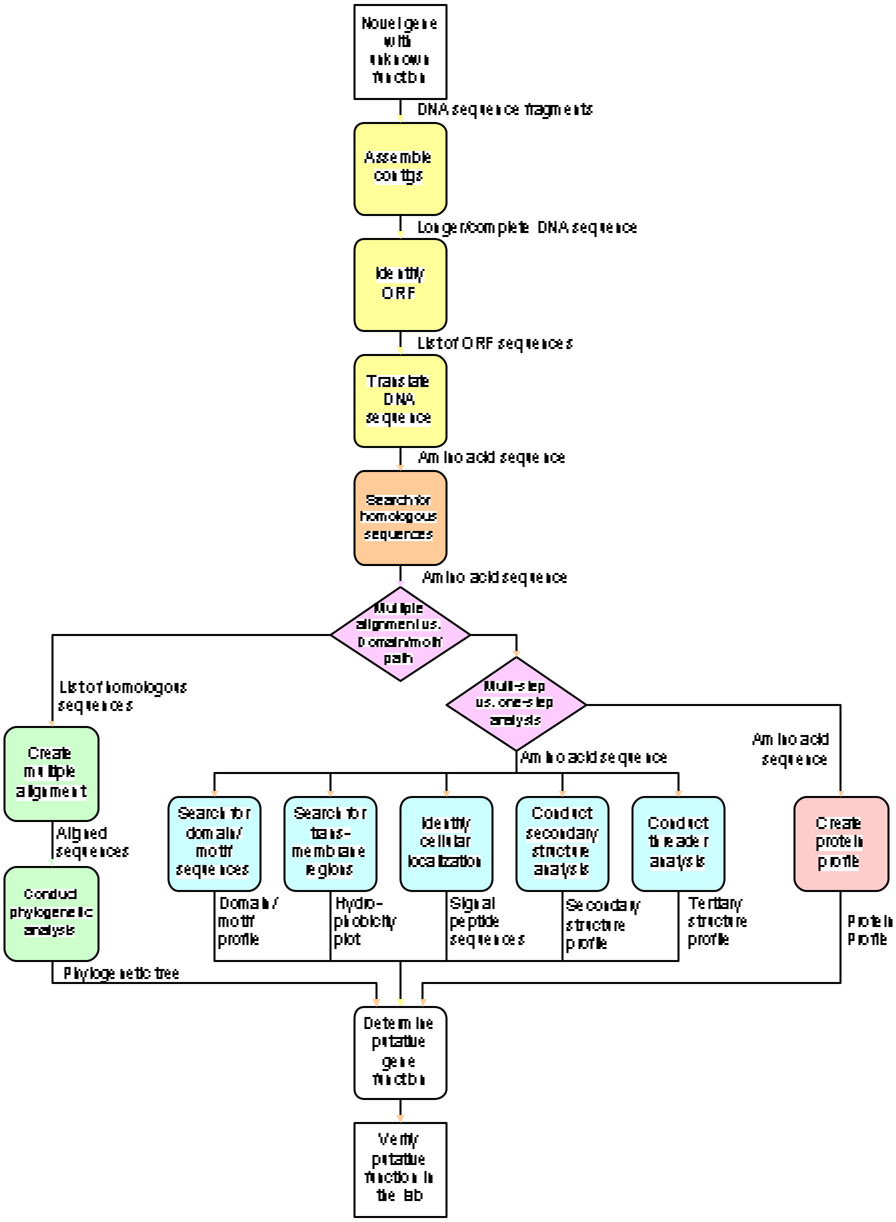

Novel gene with unknown function
DNA sequence fragments
Assemble contigs
Longer/complete DNA sequence
Identify ORF
List of ORF sequences
Translate DNA sequence
Amino acid sequence
Search for homologous sequences
Amino acid sequence
Multiple alignment vs. Domain/motif path
List of homologous sequences
Multi-step vs. one-step analysis
Amino acid sequence
Amino acid sequence
Create multiple alignment
Aligned sequences
Conduct phylogenetic analysis
Phylogenetic tree
Search for domain/motif sequences
Search for trans-membrane regions
Identify cellular localization
Conduct secondary structure analysis
Conduct threader analysis
Create protein profile
Domain/motif profile
Hydro-phobicity plot
Signal peptide sequences
Secondary structure profile
Tertiary structure profile
Protein Profile
Determine putative gene function
Verify putative function in the lab

Figure 1. Bioinformatics Analysis Protocol (High level representation)

are optional, depending on the form of the starting data. *Step 4 – Homology Searching* is the first analytical step, and is common to all three pathways. At this point, the protocol diverges into two alternate pathways, the *Multiple Alignment Path* and the *Domain/motif Path*, each of which take a different approach to the analysis of the genetic sequence. The latter further diverges into two pathways, the *Multi-Step Path* and the *One-Step Path*, both of which accomplish similar analysis, in either a single step or a series of steps.

For each step, the protocol documents:

- a definition of the step
- the rationale for including the step
- the input and output data
- how to interpret and implement the results
- what step(s) to follow next
- any caveats to consider

The protocol also describes the two key decision points, at which the process diverged between alternate pathways. Detail is provided as to why one would follow one pathway or the other.

One of the factors reflected in the protocol is that not all steps were reported by all experts, nor are they all applicable to all situations. Even in the case in which a series of steps are presented in a linear sequence, again, no one followed all of the steps. Therefore, while there is a consensus sequence and arrangement among the steps, there is also provision for the flexibility to include or skip over individual steps, depending on the scientific problem in question.

Each step of the protocol can also be seen as an information task. For each step, there is a need for information, seeking of information, and use of information.

3 Bioinformatics Analysis Protocol Interface

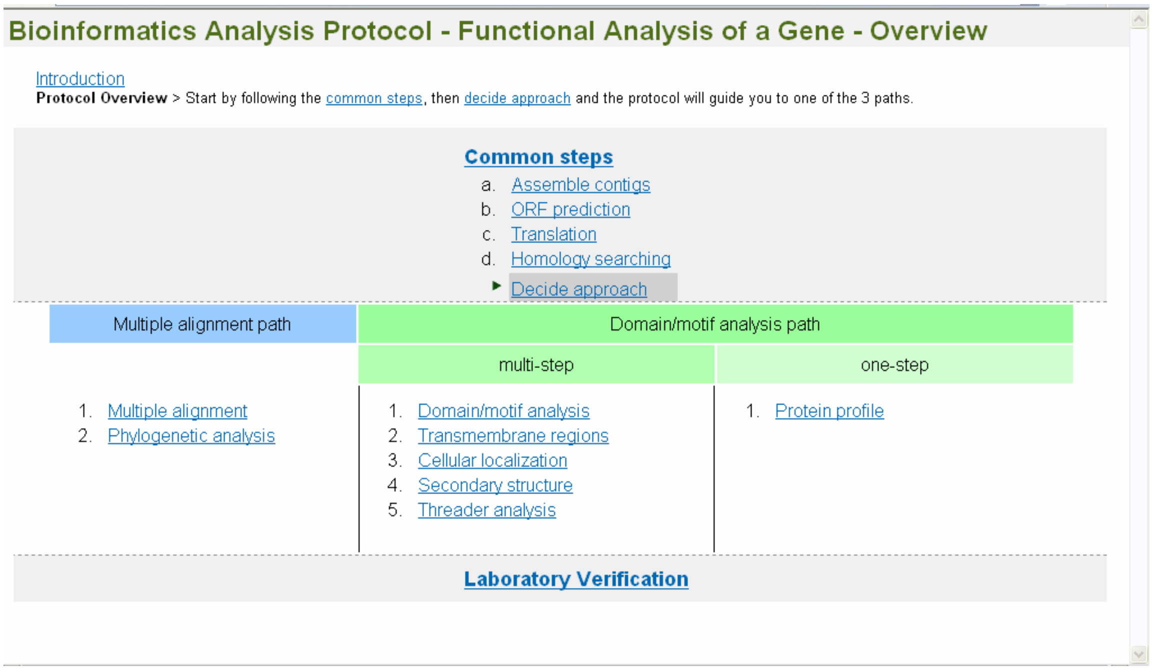
The objective of the bioinformatics analysis protocol (BAP) interface is to present the information and knowledge contained within the descriptive protocol so as to guide a laboratory scientist through the process of conducting a functional analysis of a gene sequence. The interface contains, in a formative rather than descriptive manner (Vicente 2000), all of the detailed information contained in the protocol. It also presents a roadmap through the protocol, guiding a scientist from one step to the next, while at the same time accounting for the fact that not every step will be followed. In fact, while the overview of the protocol in Figure 1 contains two decision points, each analytical step also encompasses a decision point – the determination of whether that analysis is relevant to the scientific problem. In the description of the protocol, this information was included in the detailed description for each step. In the BAP interface, it was necessary to separate the two elements of detail, that relating to the reasons for including the analysis, and those relating to how to actually carry out the analysis. Following a device-independent approach (Benyon 1992), the protocol does not include specific instructions on the use of particular bioinformatics resources.

The interface comprises a hierarchical series of web pages, paralleling the levels of detail in the protocol. At the highest level, the interface presents the protocol from a broad perspective, providing an overview and allowing orientation within both the protocol and the system. At the most detailed level there is a series of web pages, each providing description and instruction on the use and application of a specific bioinformatics analysis step. Intermediate level information supports navigation among the analytical steps, allowing scientists to determine which steps are most suitable to their particular research scenario.

3.1 Navigation within the Interface

After reading a brief introduction outlining the contents and use of the interface, a user views the protocol overview page (see Figure 2). This presents all of the steps, grouped according to the three pathways, with hyperlinks to more detailed information about each step.

There are two pages for each analytical step. The "Outline" page presents the definition of the analysis, and the rationale for why it would be included (see Figure 3). The intent is to provide enough information for the user to quickly determine whether the analysis is relevant to his or her particular situation. If a user decides to follow the step, then clicking on the "Detail" tab leads to the page containing details about how to conduct the analysis (see Figure 4). This includes links to one or more bioinformatics tools that can be used to conduct the analysis, a description of the input and output data, information on how to interpret the results (with any caveats to consider), and a link to the next step in the protocol. A user who decides instead that the step is not relevant, can then follow the link on the outline page to the next step in the sequence.



Bioinformatics Analysis Protocol - Functional Analysis of a Gene - Overview

Introduction

Protocol Overview > Start by following the common steps, then decide approach and the protocol will guide you to one of the 3 paths.

Common steps

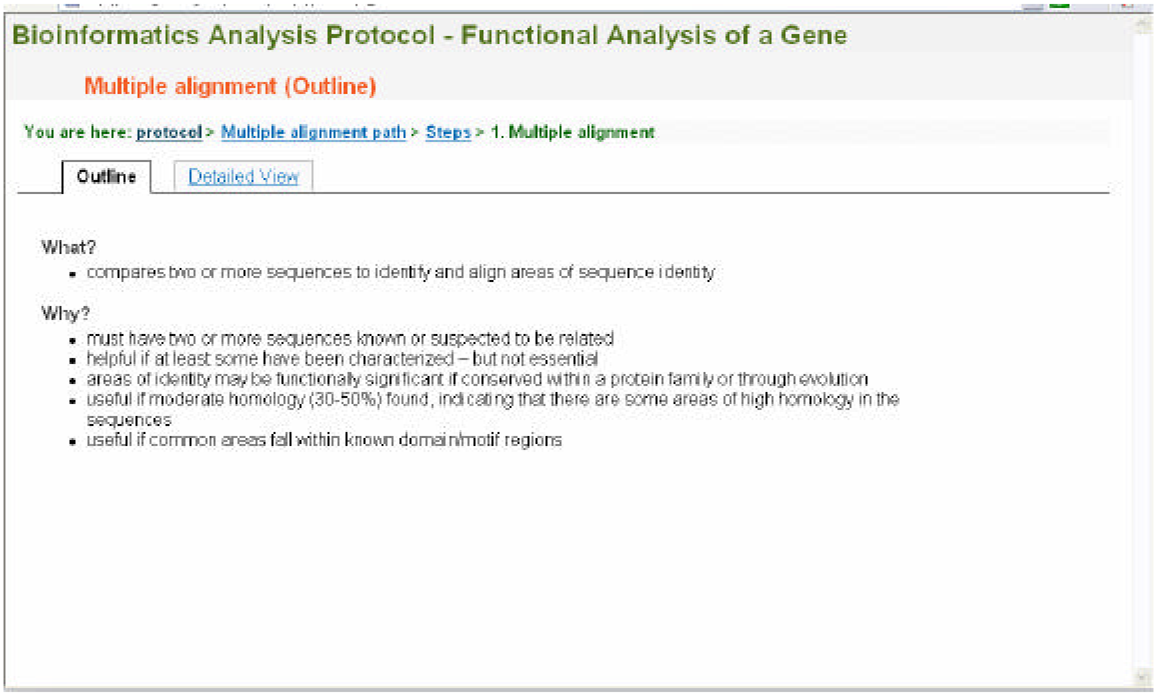
a. Assemble contigs
b. ORF prediction
c. Translation
d. Homology searching

▸ Decide approach

Multiple alignment path	Domain/motif analysis path	
	multi-step	one-step
1. Multiple alignment 2. Phylogenetic analysis	1. Domain/motif analysis 2. Transmembrane regions 3. Cellular localization 4. Secondary structure 5. Threader analysis	1. Protein profile

Laboratory Verification

Figure 2. BAP Interface Overview Page



Bioinformatics Analysis Protocol - Functional Analysis of a Gene

Multiple alignment (Outline)

You are here: protocol > Multiple alignment path > Steps > 1. Multiple alignment

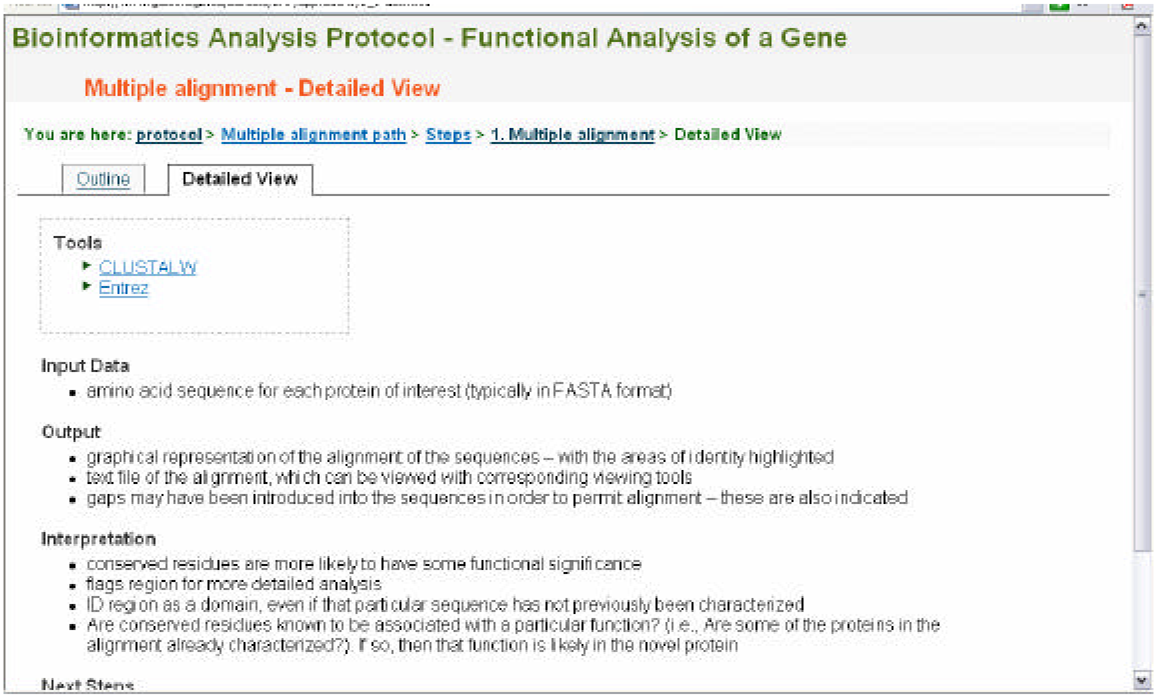
What?

- compares two or more sequences to identify and align areas of sequence identity

Why?

- must have two or more sequences known or suspected to be related
- helpful if at least some have been characterized – but not essential
- areas of identity may be functionally significant if conserved within a protein family or through evolution
- useful if moderate homology (30-50%) found, indicating that there are some areas of high homology in the sequences
- useful if common areas fall within known domain/motif regions

Figure 3. BAP Interface Outline Page



Bioinformatics Analysis Protocol - Functional Analysis of a Gene

Multiple alignment - Detailed View

You are here: protocol > Multiple alignment path > Steps > 1. Multiple alignment > Detailed View

Tools

- CLUSTALW
- Entrez

Input Data

- amino acid sequence for each protein of interest (typically in FASTA format)

Output

- graphical representation of the alignment of the sequences – with the areas of identity highlighted
- text file of the alignment, which can be viewed with corresponding viewing tools
- gaps may have been introduced into the sequences in order to permit alignment – these are also indicated

Interpretation

- conserved residues are more likely to have some functional significance
- flags region for more detailed analysis
- ID region as a domain, even if that particular sequence has not previously been characterized
- Are conserved residues known to be associated with a particular function? (i.e., Are some of the proteins in the alignment already characterized?). If so, then that function is likely in the novel protein

Next Steps

Figure 4. BAP Interface Detail Page

The two-pronged approach accommodates one of the key characteristics of the original protocol, the fact that not all steps are relevant to all scenarios. By first presenting *why* one would follow a step, and then presenting *how* to do so, users are guided to first determine whether an analysis is relevant. In this way, they can move efficiently through the protocol, selecting only those analyses that are relevant.

3.2 Implementation of the BAP Interface

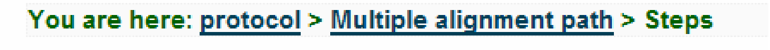
The BAP protocol interface was implemented using XHTML and styled with Cascading Style Sheets (CSS). This implementation method was chosen for its flexibility and relative simplicity.

3.3 Directory and Hyperlink Structure

There are directories for all of the common steps (including laboratory verification and decision pages) and each of the 3 approaches (*multiple alignment, one-step domain motif analysis, and multi-step domain motif analysis*.) (see Figure 5) All of the directories contain a page with an index of the multiple steps contained therein, except for *One-step Domain Motif Analysis.* The "You are here" breadcrumbs contain a link back to all pages, including the index of steps files. Each of the steps (except for *Laboratory Verification*) contains both an outline and detailed view, implemented as a tab on the user interface.

The detailed view tab of each protocol step contains a "Next steps" section that links to the next step in the approach sequence or in the case of final steps of one of the simpler approaches (*Phylogenetic analysis* or *Protein profile*) a link back up to the appropriate page of *Domain motif analysis.* (see Figure 6)

The two decision points were represented as two consecutive pages: *Decide approach - multiple alignment vs domain motif analysis* and *One-step vs Multi-step domain motif analysis.* The decision points are a part of the common steps and so reside within that directory, while the step-index pages for each of the approaches that the decision pages lead to reside within their own directories. The intention is to lead the user out of the common steps and into one of the paths (*Domain/motif Analysis, Multiple alignment*.) Breadcrumb trails are intended as an indicator as to location within the protocol, and also as a tool for backtracking to previous places in the protocol without having to use the browser back button. When viewing one of the approach steps, backtracking to the decision pages is accomplished by clicking on the path name in the breadcrumb trail. For example, if viewing *Multiple alignment (Outline)*, the following breadcrumb trail is visible:


You are here: protocol > Multiple alignment path > Steps

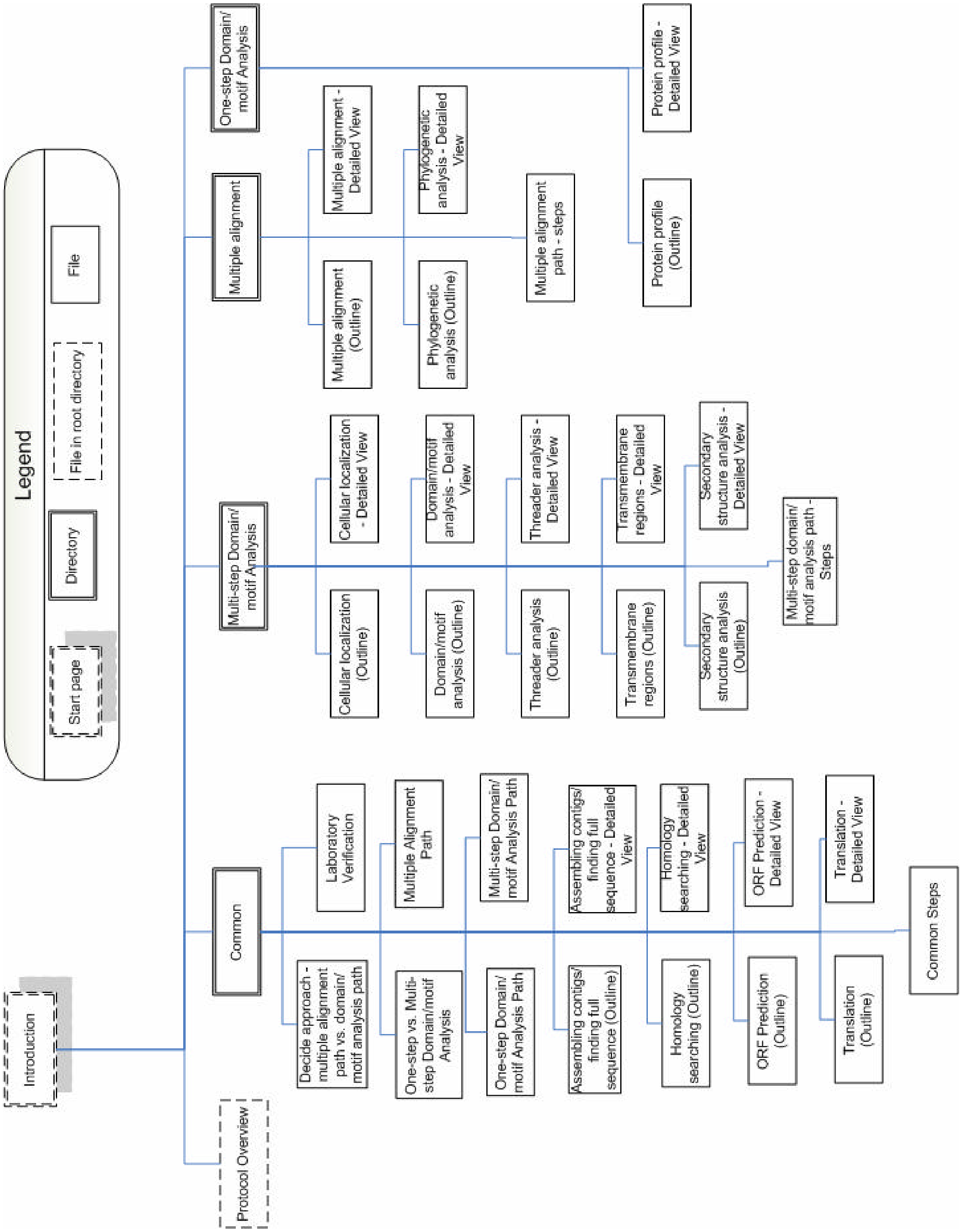
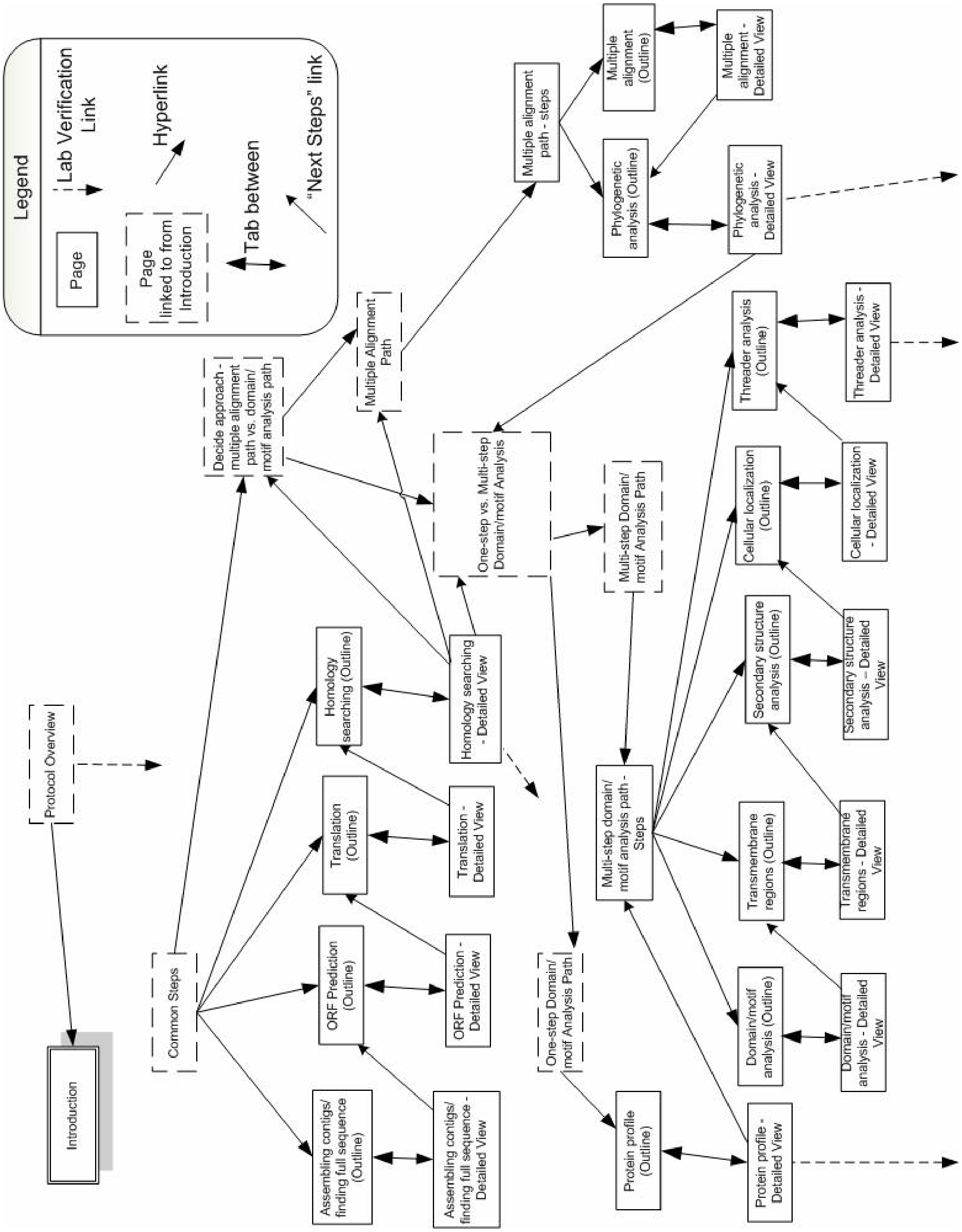

Legend
Start page
Directory
File in root directory
File
Introduction
Protocol Overview
Common
Multi-step Domain/ motif Analysis
Multiple alignment
One-step Domain/ motif Analysis
Decide approach - multiple alignment path vs. domain/ motif analysis path
Laboratory Verification
One-step vs. Multi-step Domain/motif Analysis
Multiple Alignment Path
One-step Domain/ motif Analysis Path
Multi-step Domain/ motif Analysis Path
Assembling contigs/ finding full sequence (Outline)
Assembling contigs/ finding full sequence - Detailed View
Homology searching (Outline)
Homology searching - Detailed View
ORF Prediction (Outline)
ORF Prediction - Detailed View
Translation (Outline)
Translation - Detailed View
Common Steps
Cellular localization (Outline)
Cellular localization - Detailed View
Domain/motif analysis (Outline)
Domain/motif analysis - Detailed View
Threader analysis (Outline)
Threader analysis - Detailed View
Transmembrane regions (Outline)
Transmembrane regions - Detailed View
Secondary structure analysis (Outline)
Secondary structure analysis - Detailed View
Multi-step domain/ motif analysis path - Steps
Multiple alignment (Outline)
Multiple alignment - Detailed View
Phylogenetic analysis (Outline)
Phylogenetic analysis - Detailed View
Multiple alignment path - steps
Protein profile (Outline)
Protein profile - Detailed View

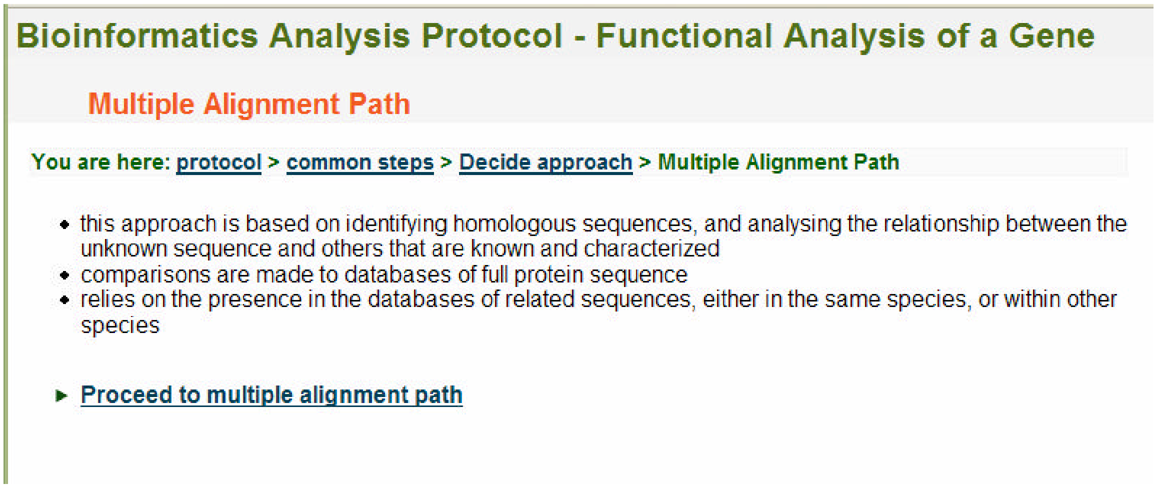
Figure 5. BAP Directory Structure


Legend
Page
Page linked to from Introduction
Tab between
Lab Verification Link
Hyperlink
"Next Steps" link
Introduction
Protocol Overview
Common Steps
Assembling contigs/ finding full sequence (Outline)
Assembling contigs/ finding full sequence - Detailed View
ORF Prediction (Outline)
ORF Prediction - Detailed View
Translation (Outline)
Translation - Detailed View
Homology searching (Outline)
Homology searching - Detailed View
Decide approach - multiple alignment path vs. domain/ motif analysis path
Multiple Alignment Path
Multiple alignment path - steps
Multiple alignment (Outline)
Multiple alignment - Detailed View
Phylogenetic analysis (Outline)
Phylogenetic analysis - Detailed View
One-step vs. Multi-step Domain/motif Analysis
Muli-step Domain/ motif Analysis Path
One-step Domain/ motif Analysis Path
Protein profile (Outline)
Protein profile - Detailed View
Multi-step domain/ motif analysis path - Steps
Domain/motif analysis (Outline)
Domain/motif analysis - Detailed View
Transmembrane regions (Outline)
Transmembrane regions - Detailed View
Secondary structure analysis (Outline)
Secondary structure analysis – Detailed View
Cellular localization (Outline)
Cellular localization - Detailed View
Threader analysis (Outline)
Threader analysis - Detailed View

Note (not shown): *Protocol Overview* links to all other pages except for the "Detailed View" tabs. Breadcrumb trail links are not shown.

Figure 6. BAP Hyperlink Structure

Backtracking to the decision pages is accomplished by clicking on *Multiple alignment path* which is linked back to the *Multiple alignment path* description in the *common step* pages. (see Figure 7)



Bioinformatics Analysis Protocol - Functional Analysis of a Gene

Multiple Alignment Path

You are here: protocol > common steps > Decide approach > Multiple Alignment Path

- this approach is based on identifying homologous sequences, and analysing the relationship between the unknown sequence and others that are known and characterized
- comparisons are made to databases of full protein sequence
- relies on the presence in the databases of related sequences, either in the same species, or within other species

▸ Proceed to multiple alignment path

Figure 7. BAP Decision Point Page

At this point, further backtracking can be done by clicking on *Decide approach* which is the entry point for the decision pages.

4. Evaluation of the BAP Interface

While we hypothesize that using the BAP interface will help a scientist to conduct the functional analysis of a gene sequence, this remains to be seen. Evaluation of the interface will involve an experimental design, with participants randomly assigned to either work with the test system (experimental group), or to follow their own procedures (control group). Each will work with a test gene sequence, and will conduct as thorough a functional analysis of the sequence as possible within a two-hour period. The key outcome we will study is the extent of the analysis of the test gene sequence obtained within a set time period. A positive result would be to see a more extensive analysis of the gene sequence obtained with the use of the BAP protocol.

We will also consider factors such as the number of bioinformatics tools used, the amount of time spent on each, the analytical steps conducted, and their order of use. It is possible that the analysis done with the use of the BAP protocol will actually take longer and encompass more steps, since participants may include more steps in their analysis -- steps that they would otherwise have been aware of. We also anticipate that the order of analytical steps will be more streamlined with the use of the BAP interface, since participants will be directed to steps in a logical sequence, rather than browsing through a variety of possibilities that might not be relevant at all, or be in the wrong place.

Participants will also complete a brief post-test survey. This will present all twelve of the analytical steps, and ask if each was used during the experiment, and why. For steps that were followed, participants will indicate whether the step was one they would typically use, or if it was new at the time of the experiment. For steps not used, participants will indicate the reason: either there wasn't time, the step wasn't relevant, or the participant wasn't sure why to include the step.

5 Conclusions

The BAP interface does not provide access to a single information resource. Instead, it integrates and coordinates the access and use of information from over seventy individual, pre-existing information systems, each of which has its own unique information architecture. The challenge was to provide a framework that logically integrated the use of a diverse variety of resources and rationally applied them to the accomplishment of a goal, and also to determine its effectiveness in supporting an information task. The user-centred design of the protocol and its interface kept the focus and emphasis on the needs, goals and objectives of the user, rather than on the system.

6 Acknowledgements

This work was supported by a McGill University New Faculty Start-up Grant to J. Bartlett.

7 References

Andrade, M. (ed.). 2003. *Bioinformatics and Genomes: Current perspectives*. Wymondham: Horizon Scientific Press.

Bartlett, Joan C. 2004. A task-based and user-centered protocol for bioinformatics information search and retrieval. *Proceedings of the ACM SIGIR'04 Workshop on Search and Discovery in Bioinformatics*.

Bartlett, Joan C. and Elaine G. Toms. 2005. Developing a protocol for bioinformatics analysis: an integrated information behavior and task analysis approach. *Journal of the American Society for Information Science and Technology* 56: 457-468.

Baxevanis, A. D. and D. B. Davison, (eds.). 2002. *Current Protocols in Bioinformatics*: Wiley [On-line].

Baxevanis, A. D. and B. F. F. Ouellette, eds. 2001. *Bioinformatics: A practical guide to the analysis of genes and proteins* (2nd ed.). New York: Wiley.

Benyon, David. 1992. The role of task analysis in systems design. *Interacting with Computers* 4: 102-123.

Butler, D. 2001. Are you ready for the revolution? *Nature* 409: 758-760.

Dalkilic, Mehmet and James Costello. 2004. BioKnOT – Biologica knowledge through ontologies and TFIDF. *Proceedings of the ACM SIGIR'04 Workshop on Search and Discovery in Bioinformatics*.

Galperin, Michael Y. 2006. The molecular biology database collection: 2006 update. *Nucleic Acids Research* 34: D3-D5.

Persidis, A. 1999. Bioinformatics. *Nature Biotechnology* 17: 828-830.

Vicente, Kim J. 2000. *Work Domain Analysis and Task Analysis: A difference that matters*. Available: http://www.mie.utoronto.ca/labs/cel.