MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Applying Task Analysis to Describe and Facilitate Bioinformatics Tasks

Dat Tran, Christopher Dubay, Paul Gorman, William Hersh

Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, OR, USA

Abstract

Objective: To document bioinformatics tasks currently performed by researchers in genomics and proteomics in an effort to recognize unmet informatics needs and challenges, identify system features that would enhance the performance of those tasks, and inform the development of new bioinformatics tools. Design: A cross-sectional study of bioinformatics tasks performed by OHSU investigators involved in genomics and proteomics research was conducted using task analysis techniques. Results: Four major categories emerged from 22 bioinformatics tasks reported by 6 research laboratories. These were: 1) gene analysis, 2) protein analysis, 3) biostatistical analysis, and 4) literature searching. Analysis of the data also raised the following challenging issues: 1) lack of procedural documentation, 2) use of home-grown strategies to accomplish goals, 3) individual needs and preferences, and 4) lack of awareness of existing bioinformatics tools. Conclusion: Task analysis was effective at documenting bioinformatics tasks performed by researchers in the fields of genomics and proteomics, at identifying potentially desirable system features and useful bioinformatics tools, and at providing a better understanding of some of the unmet needs and challenges faced by these researchers.

Keywords:

Task analysis, bioinformatics, user needs.

Introduction

In order to develop useful and usable systems, we have to understand user needs and contexts, translate them into user requirements and continually take aim at those requirements throughout the development process [1]. As Armour argued, the difficult part of building systems is not building them, but knowing what to build [2]; and it is the understanding of user needs that forms the critical basis for knowing what system to build.

One approach to understanding user needs is through a better understanding of user activities. Using this approach, Stevens et al. [3] conducted the only published study that has formally sought to capture the requirements of a system geared at supporting the informatics needs of researchers in genomics and proteomics. They conducted a questionnaire survey of 35 biologists aimed at identifying a representative collection of tasks performed by working biologists. The authors were able to classify 315 tasks into 15 broad categories (Table 1). Of these, three categories (sequence similarity searching, functional motif searching, sequence retrieval) accounted for more than half of the reported tasks.

Table 1: Classification of tasks in bioinformatics

Question class	Frequency
Sequence similarity searching	
Nucleic acid vs nucleic acid	28
Protein vs protein	39
Translated nucleic acid vs protein	6
Unspecified sequence type	29
Search for non-coding DNA	9
Functional motif searching	35
Sequence retrieval	27
Multiple sequence alignment	21
Restriction mapping	19
Secondary and tertiary structure prediction	14
Other DNA analysis including translation	14
Primer design	12
ORF analysis	11
Literature searching	10
Phylogenetic analysis	9
Protein analysis	10
Sequence assembly	8
Location of expression	7
Miscellaneous	7
Total	315

Although this study was a positive step towards understanding bioinformatics tasks carried out by biologists along with their informatics needs, it has some limitations. First, the survey was conducted in 1998. Consequently, the survey results may no longer accurately reflect bioinformatics tasks currently performed by biologists or their relative frequency, given the continuing rapid advances in genomics, proteomics and bioinformatics research. Second, interviews with domain experts formed the primary source for structural analysis of the tasks. Experts, however, are frequently not aware of all of the steps they take in performing a task—their level of experience causes some details to drop below the level of conscious thought. Moreover, the description given by an expert about a task can be misleading in that ideal rather than actual practice is provided [4, 5, 6]. For these reasons, the data for task analysis should ideally come from actual observation of the task being analyzed and interaction with experts in the field.

The overall goal of this study was to gain a better understanding of existing bioinformatics tasks undertaken by researchers in genomics and proteomics in order to better support those tasks. The specific objectives of the study were to 1) document the bioinformatics tasks and provide detailed descriptions of those that were either performed with high frequency or deemed to be crit-

ical to the laboratories' research efforts, 2) identify system features and new bioinformatics tools that would enhance the performance of those tasks, and 3) identify unmet needs faced by biologists and challenges faced by the bioinformatics community.

Materials and Methods

Design and Subjects

A cross-sectional study of bioinformatics tasks performed by Oregon Health & Science University (OHSU) investigators involved in genomics and proteomics research was conducted using task analysis techniques. Subject recruitment was based on a convenience sample selected to represent a broad range of research in these areas. The principal investigators (PI's) from the intended sample were invited to participate via e-mail. As directed by the OHSU Institutional Review Board, all subjects were asked to sign a consent form which included a description of the study protocol, any potential benefits and risks, and measures taken to maintain security, confidentiality and anonymity.

Data Collection

Data collection for each research laboratory occurred in two phases. First, a half-hour unstructured interview of the PI was conducted to provide an overview of the bioinformatics tasks that were being performed by the group and to identify the tasks that merit more detailed description and analysis. Once identified, the person primarily responsible for each task within the group was then asked to participate in the observational phase. Using the talk-aloud technique, (s)he was asked to carry out the task and verbally state what (s)he was doing without any explanation as to why particular actions were being performed. Instead, such explanations, along with questions from the researcher, were addressed in a debriefing session that followed completion of the task. This verbal protocol method was chosen to minimize the additional task-load on the person executing the task and therefore decrease the risk of interference with task performance [4, 7, 8, 9]. For each of the observed task, the researcher also asked for any available instructional guides, operational manuals, or procedural materials for the purpose of triangulation.

Data Analysis

Analysis of the data for each task occurred in two stages. First, hierarchical task analysis (HTA) was implemented, using the data from the verbal protocol session, to describe the task and provide organization and structure to its representation [10, 11, 12]. This was an iterative process involving clarification, verification and modification, as required, of the HTA output with the person performing the task until no amendments were made. Once the HTA representation of the task was finalized, the subtasks and actions were analyzed to identify features or tools that either improved task performance or provided added value. This was accomplished by relying on the researcher's insight as well as by utilizing the technique of goal composition to provide a systematic framework for analysis [13]. Once developed, the features and tools were presented to the person performing the task for validation in a categorical manner. Last, data were analyzed in an effort to identify recurring themes within the interviews and observational sessions.

Results

A call for participation in this study was sent to 9 PIs; 6 agreed to participate. Of the three who did not participate, one did not respond after 2 invitations were sent, one was on sabbatical leave and one was in the midst of training new personnel to perform the bioinformatics tasks. The research groups that did participate in the study were diverse in size, ranging from having one staff (excluding the PI) to greater than 10.

Reported Bioinformatics Tasks

Interviews of the PIs generated an initial list of 22 tasks, grouped into 4 large categories (Table 2). Attempts to group them using the classification system proposed by Stevens *et al.* [3] proved to be difficult and unsuccessful. It quickly became apparent that many of their classes of bioinformatics tasks were no longer viewed as tasks by the research groups in this study. These included sequence similarity searching, sequence retrieval, multiple sequence alignment, primer design, phylogenetic analysis and sequence assembly. This is not to say that they were no longer being performed; in fact, they continued to be common activities that were now carried out as subtasks of more complex tasks. For example, using the question classes of Stevens *et al*. [3], the task of identifying and characterizing novel nucleoside transporter protein sequences in parasitic protozoa would have included sequence similarity searching, phylogenetic analysis, and secondary and tertiary structure prediction. This evolution is not surprising and likely reflects the high level of familiarity of these tasks among the researchers such that they were now considered so simple and automatic that they no longer warranted identification as standalone tasks.

In addition, notably absent from the classification list developed by Stevens *et al.* [3] was biostatistical analysis (Table 1), which was viewed as a critical task among our participants. Yet, some of their question classes remained applicable to the tasks elicited in this study, including protein analysis and literature searching. However, compared to the study performed by Stevens *et al.* [3], where protein analysis ranked twelfth on the question class list, protein analysis was the group with the largest number of tasks in this study (Table 2). This is not unexpected given the burgeoning research efforts in the field of proteomics in this post-genomic era. After grouping appropriate tasks using these three categories (biostatistical analysis, protein analysis, literature searching), the remaining tasks were all related to genomic data and were consequently grouped into the category of gene analysis. Within the categories of gene analysis and protein analysis, there were a sufficient number of tasks that warranted further classification. Where possible, this was achieved by examining the goal of each task and grouped accordingly.

HTA of Subset of Reported Bioinformatics Tasks

From the above list of reported bioinformatics tasks, 6 underwent detailed description and analysis (Table 2) as described previously. As it is not possible to present the HTA output for

all 6 tasks, we will present here the high-level HTA result of one task and illustrate the value of this process.

Table 2: Reported bioinformatics tasks by categories

I. GENE ANALYSIS
- Identifying gene(s) underlying a disease/phenotype
  1. *Identifying and characterizing the gene(s) linked to a complex phenotype of interest*
  2. *Engineering a BAC transgenic model organism*
- Finding functional information on genes
  3. *Building an up-to-date database containing functional information on all genes of a model organism on a DNA microarray*
  4. Identifying the putative function of gene(s) of interest in a model organism based on information in human genome
  5. Searching the expressed sequence tag (EST) databases to obtain supporting evidence for information on expression of gene(s) of interest
- Identifying functionally important (conserved) motifs within genes
  6. *Searching parasite genome databases for novel NT protein sequences*
  7. Finding orthologous genes between human and a model organism
  8. Finding common motifs in genes and gene families
  9. Manipulating (i.e., introducing mutations into) sequences of genes and predicting their effects
- Other
  10. Finding genes with functional relationships (e.g., common metabolic pathways)

II. PROTEIN ANALYSIS
- Structural/functional analysis
  1. Predicting secondary and tertiary protein structure based on amino acid sequence (e.g., identifying presence of transmembrane domains) using multiple algorithms
  2. Predicting whether a protein is within or flanking some functional moiety
  3. Finding proteins that contain a domain that is structurally similar to a functional domain within a known protein of interest
- Mass spectrometry analysis
  4. Identifying unknown protein(s) in samples using mass spectrometry technology
  5. *Quantifying the relative "concentrations" of proteins in 2 samples utilizing mass spectrometry technology*
  6. Confirming a mutation site in a protein using mass spectrometry technology
- Other
  7. Localizing protein expression at cellular level
  8. *Analyzing the validity of nucleotide sequences of fusion proteins from yeast 2-hybrid screening experiments*

III. BIOSTATISTICAL ANALYSIS
1. Performing basic statistical analysis
2. Analyzing results (e.g., cluster or tree analysis) gene expression profile experiments
3. Analyzing gene-gene interaction (i.e., epistasis analysis)

IV. LITERATURE SEARCHING

Italicized tasks were those that underwent HTA

Identifying and characterizing novel nucleoside transporter (NT) proteins in parasitic protozoa

The goal here was to delineate the multigene family of NT proteins in parasitic protozoa and, in the process, gain a better understanding of functionally important regions among them. To accomplish this goal, the research group mined the emerging parasite genome databases for protein sequences that are homologous to known NT proteins.

As one can see from Figure 1, completion of this task involved multiple iterations, one for each combination of every known NT protein sequence against every available parasite genome database. At the time of this study, there were 22 parasite genome databases and 9 known NT proteins, although this number is likely to increase. However, due to time constraints, the group was only searching on 2-3 of the known NT proteins against 2-3 of the parasite genome databases. Even with this limited search strategy, it usually took half a day to complete. Therefore, a script written to link all possible combinations of searches would have been enormously useful. Alternatively, time savings could have been accomplished with an algorithm that produces a degenerate sequence from the sequences of all known NT proteins that then could be used to search all of the parasite genome databases. For this to be effective, the algorithm must allow the user to stipulate known conserved motifs with specified order and spacing, in addition to being capable of identifying potentially novel conserved residues, in producing the degenerate sequence.

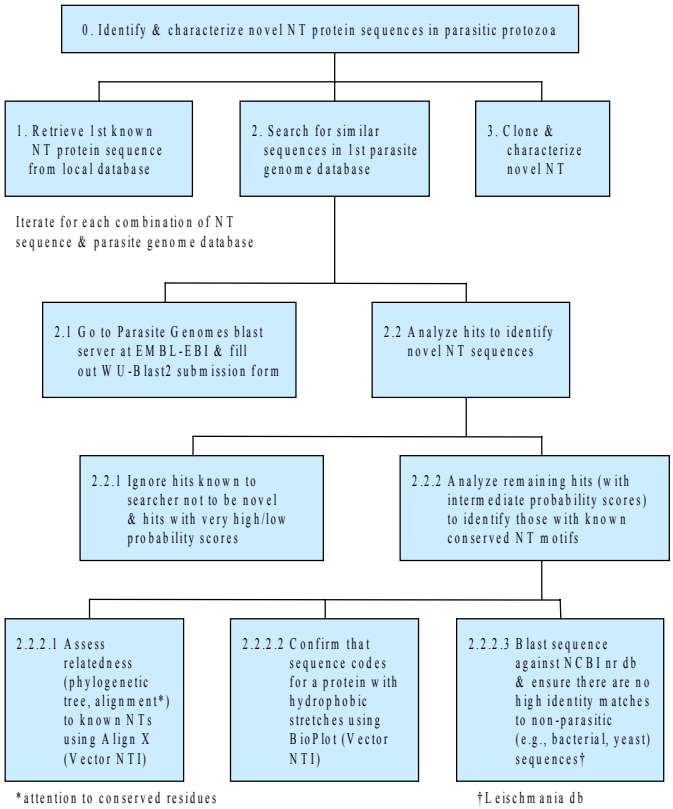

Figure 1 - HTA diagram of the task of identifying and characterizing novel NT proteins in parasitic protozoa.

At the time of this study, identification of new sequences relied on the searcher's memory. When searches are not performed frequently, as it was in this case (once a month), such an approach could lead to unnecessary review of hits that have already been analyzed in previous searches. This could be avoided by maintaining a database of cumulative results from previous searches to be used in filtering out old hits. The database could

be implemented either on the client side or on the server side in the form of an account. A more efficient strategy would be to tag each sequence in the parasite genome databases with an addition/update date, which would allow the searcher to limit retrieval to sequences that had been added or updated after a specified date (presumably the last search date). Another approach to minimizing the number of hits returned by a query would be to equip the current BLAST algorithm with the ability to require mandatory matching to specified (conserved) motifs for a sequence to be returned as a hit.

Themes

Further analysis of the interviews and verbal protocols raised a number of thematic issues, which are presented below.

Lack of Procedural Documentation

None of the laboratories that participated in the study had a written protocol, not even a high level one, for any of the bioinformatics tasks that were being performed.

Use of Home-Grown Strategies

The approaches used in all of the observed tasks in this study were developed in-house, with no attempt to seek alternative strategies used by other laboratories performing similar tasks.

Individualized Needs and Preferences

Despite the need to acquire tools that would achieve very similar goals, there was very little overlap in the applications that were used by the 6 participating laboratories. Some of this could be attributed to the concept of individualized needs. For example, in task 15 (Table 2), the particular needs of the laboratory precluded the use of already available applications and required the development of an in-house application. However, this lack of overlap was not only observed among research groups but also among individuals within each research group performing very similar tasks. Clearly, this also speaks to the importance of individual preferences in tool and application selection.

Lack of Awareness of Existing Bioinformatics Tools

The following is a representative quotation that highlights this issue: "…I'm sure there are already tools out there that can do this more efficiently, I just don't know what they are or even where to begin looking." A tangible instance of this problem arose in the execution of task 2 (Table 2), where even though the task could have been performed more economically with an existing application, it was not, and all due to the fact that the user was not aware of its existence.

Discussion

Themes and their Implications

The issues, described above, in turn, present challenges for not only biomedical researchers but also for the bioinformatics community in its effort to support these researchers' goals.

Impact of the Lack of Procedural Documentation

The pervasive lack of procedural documentation presents several challenges. One is the issue of reproducibility of the task, especially in cases where a task is performed by a number of staff members. But even when it is performed by one person, day-to-day variability in carrying out that task may exist as a result of a number of factors, such as variation in time pressures, level of fatigue or well-being and the day of the week. This brings into question the reliability and quality of the data retrieved without written protocols. The lack of documentation also makes it difficult for other investigators who wish to validate the results by applying the same strategy. Last, in this climate of short supply of personnel with bioinformatics skill-set, the departure of the person primarily responsible for the performance the task(s) would make training more difficult and cause more disruption in the absence of documented protocols.

Impact of the Use of Home-Grown Strategies and Individualized Needs and Preferences

Given that strategies used for all of the tasks in this study were developed in-house with little comparative evaluation, they are potentially sub-optimal. Moreover, their pervasive existence, combined with the matter of individual needs and preferences, requires development of customized tools and precludes the possibility of automating and supporting such tasks in a generalized manner. Yet, the majority of biomedical researchers do not have the resources to support an in-house person dedicated to developing customized bioinformatics tools, nor do they have the monetary power that health care organizations have in the electronic health record industry to demand customizable applications from commercial vendors. This dilemma is, to a certain extent, offset by the wide array of available bioinformatics applications and tools. This, however, leads to an almost equally diverse range of such tools and applications in use amongst researchers. The multiple data formats that result from the use of different applications by collaborators present a significant barrier to data access and sharing, as well as the development of common tools aimed at manipulating the collaborative data.

Impact of the Lack of Awareness of Existing Bioinformatics Tools

This study illustrated that while existing bioinformatics tools/applications are available to better support the bioinformatics tasks that are currently performed, they may be effectively inaccessible because of the user's lack of awareness of their existence. This issue underscores the great challenge faced by the bioinformatics community to not only develop tools, but also to effectively deliver and disseminate them.

Limitations

This study has several limitations. The data were collected in a single academic institution from a small sample of research laboratories as well as from observation of a small number of bioinformatics tasks. As such, the findings may not be generalizable beyond the participating laboratories and may not apply to an industry setting. In addition, the results were based on observation and analysis of one instance of each task, and that instance may not have been a representative one. Further, even though an attempt was made to minimize interference with task performance, it is still possible that the verbal protocol had an intrusive effect on and therefore altered the usual procedural steps taken to complete the observed task. Finally, selection bias might also be a concern in that those laboratories that agreed to participate in the study were more likely to face challenges and

have unmet needs. However, it was explicitly communicated to the PIs that any potentially useful features or tools identified from this study would not be implemented as part of the study.

Possible Future Directions

Given the small sample of research laboratories as well as tasks examined in this study, a larger study aimed at analyzing a broader range of tasks in other laboratories, both at this institution and at others, would improve the generalizability of findings and may uncover other important issues faced by researchers in genomics and proteomics. With a large enough sample of tasks, a more refined and robust classification system could also be developed and validated using a formal methodology such as that used by Ely *et al.* [14]. Another natural extension of this study would be the implementation of some of the proposed features and tools and validation of their value with objective measures. Some of the challenges identified in this study also present numerous opportunities for future research. Assuming that the tasks observed in this study are also performed elsewhere, development of generalizable tools and systems to support these researchers would be ideal. However, this would only be feasible if there were at most a few generally accepted approaches to completing a particular task. Hence, a potentially useful study would be one that attempts to develop and evaluate such "best practice" protocols. However, given that the need for customization will always be present, there is also the opportunity to develop, implement and evaluate different models of support services to meet individual needs and preferences.

Conclusion

Task analysis was effective at providing a low-level description of some of the bioinformatics tasks performed by researchers in the fields of genomics and proteomics and at identifying potentially desirable system features and useful bioinformatics tools. Moreover, it provided a better understanding of some of the unmet needs and challenges faced by these researchers and the bioinformatics community, with many of those challenges being related to lack of standardization of procedures and protocols. More research is needed to validate and generalize these preliminary findings.

Acknowledgments

This work was supported in part by the National Library of Medicine through a Biomedical Information Science and Technology Initiative (BISTI) Administrative Supplement to the OHSU Fellowship program in Medical Informatics (Grant #:T15 LM07088).

References

[1] Beyer H and Holtzblatt K. *Contextual Design: Defining Customer-Centered Systems*. San Francisco: Morgan Kaufmann Publishers, 1998.

[2] Armour PG. The Five Orders of Ignorance. *Communications of the ACM* 2000: 43 (10): 17-20.

[3] Stevens R, Goble C, Baker P, and Brass A. A Classification of Tasks in Bioinformatics. *Bioinformatics* 2001: 17 (2): 180-188.

[4] Kirwan B and Ainsworth LK. *A Guide to Task Analysis*. London: Taylor and Francis, 1992.

[5] Carey MS, Stammers RB, and Astley JA. Human-Computer Interaction Design: The Potential and Pitfalls of Hierarchical Task Analysis. In: Diaper D, editor. *Task Analysis for Human-Computer Interaction.* Chichester, UK: Ellis Horwood, 1989. p. 56-74.

[6] Lewis C. A Research Agenda for the Nineties in Human-Computer Interaction. *Human-Computer Interaction* 1990: 5 (2): 125-143.

[7] Ericsson KA and Simon HA. *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: MIT Press, 1993.

[8] Weidenbeck S, Lampert R, and Scholtz J. Using Protocol Analysis to Study the User Interface. *Bulletin of the American Society for Information Science* 1989: 15 (5): 25-26.

[9] Poulson DF, Ashby MC, and Richardson SJ. USERfit: *A Practical Handbook on User-Centred Design for Assistive Technology.* Brussels, Luxembourg: ECSC-EC-EAEC, 1996.

[10] Shepherd A. HTA as a Framework for Task Analysis. *Ergonomics* 1998: 41 (11): 1537-1552.

[11] Dix A, Finlay J, Abowd G, and Beale R. In *Human-Computer Interaction.* Hemel Hempstead, Hertfordshire: Prentice Hall Europe, 1998.

[12] Preece J, Rogers Y, Sharp H. *Human-Computer Interaction*. Reading, MA: Addison Wesley, 1998.

[13] Nielsen J. Goal Composition: Extending Task Analysis to Predict Things People May Want to Do. 1994. http://www.useit.com/papers/goalcomposition.html (20 July 2002).

[14] Ely EW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, and Stavri PZ. A Taxonomy of Generic Clinical Questions: Classification Study. *BMJ* 2000: 321: 429-432.

Address for correspondence

Dr. Dat Tran, Oregon Health & Science University, Department of Medical Informatics & Clinical Epidemiology, Portland, Oregon 97239-3098 (tranda@ohsu.edu).