BIO-150
SAC2005
Design and Evaluation of CATPA: Curation Alignment Tool for Protein Analysis

September 2, 2004

Abstract

We present a new application for experimental biologists, the Curation Alignment Tool for Protein Analysis (CATPA), that allows for the efficient and effective creation, storage, management, and querying of experimentally curated protein families. As the number of discovered genomic and proteomic sequences outpaces our ability to understand them, the experimental biologist, who is our primary link in fundamentally and essentially understanding genomic and proteomic information, is left further behind in our race to automate and semi-automate information discovery. With this in mind we have designed a system to help them "catch-up" as it were with the systems devoted to annotation. In order to help evaluate the utility of this system, we have performed a formal quantitative and qualitative usability study. Usability evaluation of CATPA was performed to compare the performance of CATPA with a similar tool PFAAT, and users were found to to significantly more efficient using CATPA for a number of different types of tasks.

1 Introduction

It is a widely accepted conjecture that by understanding the geometry of a protein, its function will likewise be known [28]. Although a protein's three-dimensional structure is determined from its amino acid sequence, reckoning this structure *ab initio* solely from its sequence information has proven quite challenging. Furthermore, once a structure has been elucidated, deciding what role the protein plays is still as difficult. One of the more successful ways biologists have dealt with this problem is by observing the maxim *like behaves like*. Presumably, if sequences are similar enough, then collections of these "like" proteins with associated information–*families*–can be used to piece together an understanding of their functions [19, pages 93-112]. Dayhoff originally described the notion of family: proteins of like biochemical function that share at least half of their sequences when aligned [12]. Interest is drawn naturally to regions of *conserved* residues, or *motifs*, that represent both the starting point and ending point of figuring out what a protein does.

As a starting point, a motif may contain little information beyond conservation, or be simply a biological hunch. Biologists then experimentally modify sets of residues contained in the motif while observing how properties of the proteins, or *phenotypes*, change. For example, ligand binding might be attenuated in a nuclear hormone receptor [17]. By conducting these experiments, information about the motif is gradually filled-in culminating in an understanding of what the proteins' functions are. This process of human-directed discovery is called

curation. In contrast, because of the rapidly increasing number of sequences available, most motif information is discovered computationally, generally via some probabilistic model. This process of automated discovery is called annotation. Each of these two processes has benefits and drawbacks. Curation, because of its meticulousness, is slow, but the information is significant and correct. Annotation, while being much faster, is error-prone and generally shallow [20, 9, 14, 10]. Again, because of the increasing backlog of sequences, research has tended to focus on public systems and repositories aimed at annotation, while curated systems have lagged far behind. Clearly what is needed is a means of combining these two processes, but a number of hurdles must be overcome.

First and foremost biologists require a means of effectively managing and querying protein families and curations. Secondly, biologists need a means of interacting with their data. Biologists have historically interacted with proteins visually; consequently, for the system to be successful, there must be an equivalent mechanism available to interact with their data. It is an exciting prospect that there may likely be more families to discover than have already been discovered [24].

CATPA, Curation Alignment Tool for Protein Analysis, is an information system for biologists that creates, stores, manages, and queries curated protein families. CATPA utilizes a GUI front-end that provides biologists with an environment they are accustomed to, while at the same time, providing the integrity and power of a local, stand-alone database at the backend. CATPA incorporates the GO ontology [18] in its curation vocabulary. In this paper we discuss the design and implementation of CATPA and a quantitative evaluation of its design.

The outline of the paper is as follows. We will first review the state of related bioinformatic software applications in Section 2. We then introduce the concepts of curation and annotation in Section 3. Section 4 introduces the CATPA system, and Section 5 elaborates on its design. In Section 6, we discuss methodologies and results from an extensive usability evaluation study using both quantitative and qualitative techniques. Finally we present a conclusion and summary of this work in Section 7.

2 Bioinformatic Systems

Dayhoff was one of the first to create a database of proteins–publishing an "Atlas" of protein sequence and structural information [12]. This initiated a coherent structure that could, in part, be used to share information. In this section we will highlight the current state of affairs in biological databases. We will then highlight a number of different systems. We will then conclude by focusing on a single, widely utilized example of a protein family database, the PIR (Protein Information Resource).

The problem of representing, managing, querying, and sharing biological data has only begun to have been addressed in the last couple of decades [6, 4], though most of the focus has been upon sequences *i.e.*, DNA, RNA, and protein. Various levels of annotation and organization help make this information richer. Indeed, new information is being generated simply from human observation of the composite information that is accumulating about these sequences. Though DNA, RNA, and proteins are quite similar from a number of perspectives *e.g.*, encoding their structure as strings, historically the means of how biologists began representing, storing (format), and so forth, has lead to generally separate paths for their development. There is a recognition, however, that all these data must eventually be integrated. Furthermore, all biological information, eventually from every aspect of every organism, must

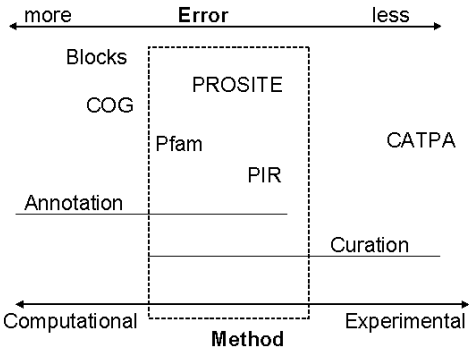

Figure 1: The spectrum of adding information from curation to annotation.

be available for querying, corroboration, *etc.*. An excellent feel for the magnitude of this task is the annual issue of Nucleic Acids Research that lists several hundred biological databases [4]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [23] is an excellent example of integration. This database endeavors to capture metabolic pathways for completely sequenced genomes. A fuzzy, but useful, demarcation of these databases is based upon whether it is (1) simply a repository for sequences; (2) an annotation (or curation) of sequences; (3) a complex collection of information associated with sequences and their annotation and curation. Well-known repositories are PDB (Protein Data Bank) [34], EMBL (European Molecular Biology Laboratory) [31], GenBank [5], Pfam [3] and PROSITE [30, 16]. Annotations can be found in Swiss-Prot [7]. Flybase [32] at Indiana University has long been recognized as a leader in this type of system.

Generally these systems lie along another fuzzy dimension of annotation where at one end, information is discovered via computation, and at the other humans, using their domain knowledge and access to other materials–generally electronic–hand annotate the sequence information. Of course, experimentally determined information–which must be curated–generally is drowned out from the excess of annotated information. Furthermore, tracking how the information arose–whether from machine, human annotation, or experimentation, is generally ignored.

As mentioned in the introduction, understanding a protein's function solely from its sequence information is yet impossible. Biologists have dealt with this problem by creating families–sets of "similar" proteins. These sets, in general, do not partition a set of proteins in the formal sense. Indeed, a protein can and usually does belong to several families. A number of different kinds of approaches in how families are formed and kinds of information associated with them has arisen within the last several years. Furthermore, families may include only proteins from a single organism–paralogous–or from multiple organisms–orthologous. We will simply highlight a number of different systems here, but the reader will find an excellent presentation in [36, 35, 7]. There are families based on domains, *e.g.* Pfam [3]. Sequence motifs are found in PROSITE [16]. Curated systems together with automated annotation is seen in Swiss-Prot and TrEMBL [7]. Families can be established via structure, *e.g.* SCOP [11].

2.1 PIR and the PSD

It is instructive to briefly examine one of the most widely utilized protein family annotation systems because (1) demonstrates what kind of data biologists are interested in keeping track

of; (2) gives an example of what kind of interface biologists are used to; (3) provides motivation and ideas for kinds of functionality a curation system should have.

As mentioned in the introduction, most recent work has been devoted to public systems for annotating protein families. One of the most widely utilized systems is the Protein Information Resource (PIR) [35, 36]. This system is a public resource for protein informatics providing, amongst a number of services and information, and functional annotations. From PIR, biologists can search the Protein Sequence Database (PSD), a collection of functionally annotated protein sequences. The genesis of PSD was from Dayhoff's Atlas of Protein Sequence and Structure which began in the mid 1960s.

Currently the PIR possesses nearly 300K entries. One distinguishing feature of the PIR is the superfamily classification: a hierarchical, non-overlapping structuring of proteins that likely reflects their evolutionary relationships.

2.2 PFAAT

The Protein Family Alignment Annotation Tool (PFAAT) [21] is a tool written in Java, which is designed to facilitate the analysis and annotation of large protein sequence families. PFAAT allows users to perform many pertinent tasks such as aligning collections of sequences, clustering and grouping sequences into subfamilies, and for analyzing sequences based on similarity. PFAAT can also aid in visualizing protein structure, and allow users to annotate sequences and specific residue positions with textual descriptions.

Given the similarity in the basic operations of PFAAT with CATPA, we have identified PFAAT as a target system for comparative analyses with CATPA. In Section 6 we will present a comparative user study analyzing the efficiency and accuracy of users using these two systems.

3 Curation vs. Annotation

In reviewing these sequence and annotation databases, it is very easy to lose sight of the reason why they exist–to facilitate *humans* to understand biology. Paradoxically, little attention has been paid to the individual biologist or lab experimentally discovering information about sequences. In our case, we are particularly interested in protein function. Figure 1 illustrates the spectrum of research conducted. At one end, the biologist experimentally determines the nature of the protein. This is *curation* in it purest sense. At the other end is computationally determined information. This is generally derived from probabilistic models, *e.g.* hidden Markov Models [19, pages 77-91]. This is *annotation*. These two means of generating information are not at odds. Indeed, in between lies a mixture of annotation and curation. Because of the ever growing plethora of available information, humans can "curate" annotated data *i.e.*, an expert can recognize and combine information drawn from both sides. As mentioned previously, human curation is slow at best. Because of this, research has focused on the annotation side of these databases.

3.1 Curation via Experimentation

Because our contribution lies on the curation side, it is instructive to discuss what a biologist who is working with a protein family does experimentally. First a biologist needs to identify motifs–active, important regions of a protein that are likely shared among most, if not all, of the proteins in the family. Discovering motifs is a difficult task. Generally recognition of motifs are either due to wise and experienced speculation by the biologist or are suggested

computationally as they appear to be, for example, statistically significant [33, 1], or through combinatorial pattern discovery [27, 22]. Motifs often are suggested visually by regions of similar residues. Proteins are aligned to achieve some optimal match. Regions of particularly strong matches are said to be *conserved regions*. Individual positions in different sequences are associated with either each other or *gaps* (non-residues) which allow a kind of stretching of proteins to achieve a more optimal match. Interestingly, alignments are significant because they can strongly suggest motifs.

Matches themselves rely upon scoring matrices [25]. These are collections of ternary relations that indicate the strength of similarity among residues which, in turn, determine whether a match (and therefore alignment) is optimal. There are many different similarity matrices that are available, their choice being determined by the context of the protein family, the biologists' experience, and so on. Once a candidate motif is identified, the biologist then can experimentally determine not only whether the region is actually a motif–that is, biologically significant–but also the very nature of the motif. This determination is done by changing generally one or more residues and observing either a change in existing phenotypes or the presence of new phenotypes. By "change", we mean a residue is either completely replaced by another residue–say a hydrophobic by a hydrophilic–or the residue itself is modified, perhaps being methylated. In a motif then, replacing a single residue, for example, can attenuate the ability of a protein to bind to its substrate. In a sense, the biologist is filling in information about the motif. Typically, information is associated with these sets as a whole–in addition to information about the individual residues. Obviously, it can be the case that the biologist discovers that the region is not biologically important.

Clearly, this kind of research is essential to our complete understanding of protein families and their respective members. What is problematic for the experimental biologists is that they have been largely ignored and forgotten in the race to provide better bioinformatics for annotation.

4 A Curation System

In summary then, we are able and have been generating an ever increasing number of sequences and are being outpaced by our inability to make sense of it. Consequently, we have focused on creating systems that are (1) public that allow all biologists to contribute information, hopefully, increasing the likelihood that information will eventually be added by someone (2) annotative that allow us to computationally suggest information that will guide biologists to significant and pertinent information (3) inclusive of annotative as well as curative information. Fundamentally driving all of this work, furthermore, is the experimental biologist who shares a keen interest in discovering biological information–in our case, protein function–but who is disconnected technologically from these systems, further exacerbating the process of bringing curation information to the fold.

We propose a bioinformatics system for the experimental biologist that allows for the efficient and effective creating, storage, management of curation information for protein families. Like its annotative counter part, the system should have the following capabilities: The system should rely on recognized formats; The system should have a GUI (graphical user interface) that allows for most of the typical tasks the biologists performs with respect to visualizing alignments, curated residues, *etc.*; The system should possess a rigorous data model and consequently a DBMS to handle the management and security of the stored information; The system should include a querying capability that allows efficient and effective querying of the

entities in the database; The system should have a means of standardizing the vocabulary so that biologists can more easily and correctly search and share information; The system should function as a standalone entity which is better suited and can be tailored to the biologists' needs; The system should possess the facility for querying and sharing information between two users if both agreed upon and desired.

Although a detailed description of a number of key elements of CATPA will be presented in the following section, we will highlight them here. CATPA recognizes a number of well-known and widely used formats both for importing to and exporting from the system. CATPA utilizes a Java GUI front-end that allows biologists to interact with information in an environment they are accustomed to. Protein families are aligned, conservation and curation are easily discernable via colors that users can change according to their preferences. Additionally, other kinds of information can be displayed, *e.g.*, entropy, hydrophobicity.

CATPA provides two separate views of the protein family. One is the standard view of an aligned family. The other view is a facility to magnify (increasing or decreasing) over the family called the Dataset View. In addition the Dataset View is used to visualize the results of queries over the protein family. CATPA has a data model and a generated logical schema. MySQL runs at the backend. All the information in the system, from alignment, to protein, to curation is modeled relationally and is created, managed, and queried via the data definition language (DDL), data manipulation language (DML), and structured query language (SQL). CATPA has extensive query facilities including the ability to query alignments, curations, sequences, and fixed vocabulary. Additionally, CATPA allows for the visualization of query results making perusal easier. To help standardize vocabulary, CATPA utilizes the Gene Ontology classification scheme (GO) [18]. GO is a collaborative effort to help standardize biological words by providing generalization and component relationships ("is-a" and "is-a-part-of").

CATPA has a small footprint, comprising less than 10MB. The front-end requires Java 1.4.2 and MySQL and can be run on most machines capable of running both. Although not completed in the current release, CATPA is intended as a means of collaboration–biologists wishing to share and query other CATPA databases can do so transparently.

In the next section we will discuss the overall design of the system. This provides an interesting glimpse into the engineering aspect of CATPA. Following that we will discuss in more detail several elements–visualization, curation, and querying facilities. We then will discuss an extensive user-evaluation involving both quantitative and qualitative analysis comparing CATPA with PFAAT, a protein annotation system discussed earlier.

5 CATPA: Design and Engineering

Software engineering is a mature, though still changing, discipline that aims to formalize the task of designing software. Software engineering is a difficult prospect at best. This is due mostly to unclear, changing, or lack of specifications. One particularly challenging problem is separating the specification from implementation. Too often "what to do" is confused with "how to do it". This is especially true in bioinformatics where there is a general paucity of even a discussion of software engineering principles. Furthermore, since most biologists prefer to work visually with their data, pinning down exactly what is specification versus what is decoration is problematic.

In designing CATPA we began by consciously ignoring the visual and focused on the "what to do". It became clear that there were really only a few kinds of entities involved and that

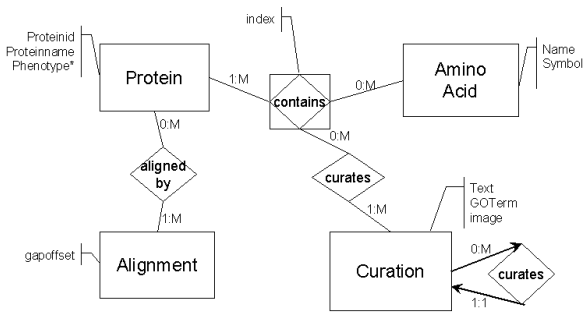

Figure 2: The ER Model for the CATPA architecture

the different terms used by biologists were simply different instances of a more general class. Furthermore, though there were many apparently disparate terms, biologists were interested mainly in answering questions about what proteins and, more specifically, what residues had curations, what was the content of the curations, and what alignments were used. We observed that there are sequences (proteins) and amino acids (residues) that comprise the sequence. Clearly, one or more pieces of information can be associated with a protein, likely an observed phenotype. Information (one or more) can be associated with a residue. Collection of these pairs, information and residue, can be themselves associated with information. Alignments are n-ary relations among proteins (with gaps). Interestingly, all of the biological information– from motif, to curation, to "super-curation" (a curation of curations) are modeled from the above. The notion that a motif and curation really are the same kind of thing would be untenable to biologist, but to a software engineer, it makes perfect design sense. We then began discovering the visual elements required to effectively manage and query this model.

We were sensitive to visual design principles *i.e.*, human computer interaction (HCI), using mental models and principles of familiarity [26]. One result was the creation of a visualization for the results of a query. Consistently we observed that biologists were hungry for discovery about their own data. They did not, however, want to view text. Our solution was to provide a database view together with a lower magnification and points indicating where the query was true. The query facility is rather flexible and allows for searching at the residue, protein, or alignment level. Obviously included is the ability to search for the presence or absence of terms in the curation. In keeping with the general theme of "freeware" in bioinformatics, we decided early on to use a GUI and database that were freely available and widely used.

In Figure 2, we illustrate a general Entity-Relationship (ER) model for the CATPA. An interesting observation is the promotion of alignment to entity status.

In the following section we focus on several design elements of CATPA by walking through a simple query over a family.

5.1 Specific Design Elements of CATPA

Figure 3 shows the main CATPA user interface. There are three main areas. The topmost area is the standard menu which allows for standard tasks some of which can be done graphically. The middle area is called the *workplace*. In the workplace, there are several tabs that allow the user to isolate sets of proteins through different kinds of queries. The proteins can be shown explicitly with residue, index information or graphically *e.g.*, conserved motif information. The bottom area houses an alignment of proteins. In this area the user can search and edit proteins, alignments, motifs, and curations. Different colors–all user definable–denote

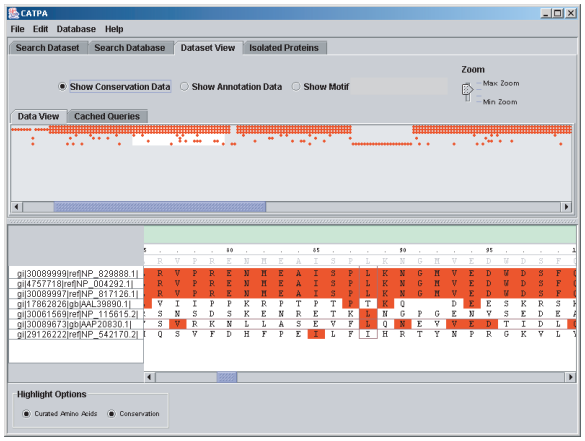

Figure 3: The CATPA user interface. The alignment shows conserved regions (shaded). The workplace is a lower magnification of the alignment. The portion of the alignment that is visible is the white rectangle in the workplace.

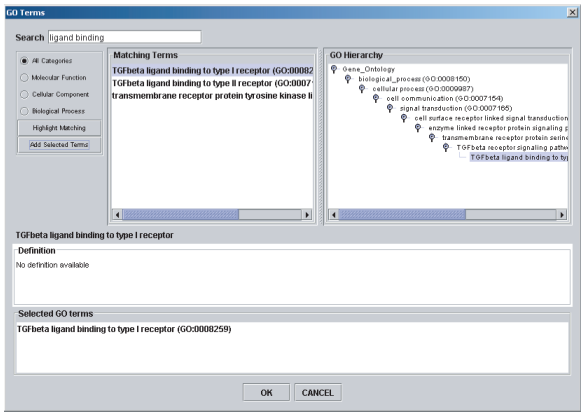

Figure 4: CATPA Interface: Browsing GO terms

different information. Usually, experimental biologists are interested in motif information, conservation, and where and what curations occur. These are denoted by coloring the cells above the residues, for instance in the motifs, or the cell itself, for example curations. As the user passes over a curation, the set of GO terms associated with the curation is displayed as a list at the bottom of the application.

We will next walk through a simply query in order to view some of these elements in action. Suppose the user is interested in locating any curation involving ligand binding type I receptor for a glutamine residue. The user goes to the workplace area and selects the Search Dataset tab. In the Amino Acids box, the user enters Q. Then the user clicks the Choose Go Terms button and database search screen is opened that allows for easy searching of GO terms. The user enters "ligand binding" in the Search box, selects "All categories" radio button and hits enter. All the GO terms containing ligand binding are listed together with their respective ontologies for the biologist to peruse. Terms can then be added to the Search box by selecting the GO terms the biologist is interested in–in this case, the type I receptor is the first term in the list (See Figure 4).

The user then issues the query against the alignment, though the query is issued actually

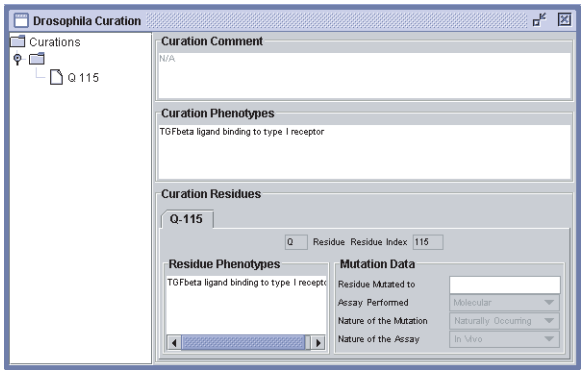

Figure 5: CATPA Interface: A curation

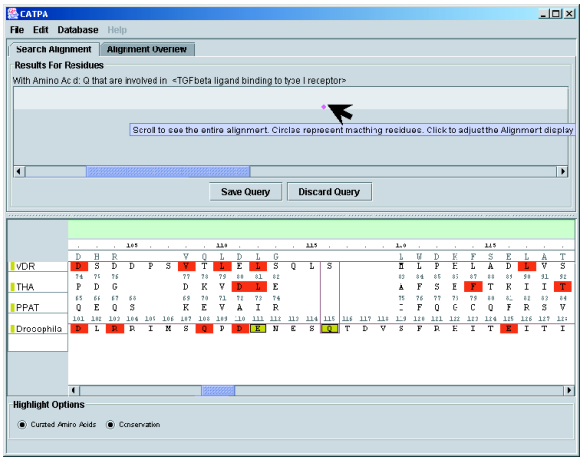

Figure 6: CATPA Interface: Querying results in CATPA. The workplace shows a point (with the mouse arrow) that when clicked moves to the residue in the alignment below.

against the database and the visualization is produced that shows as a series of points where the residues (if any) reside. The user scrolls through the dataset in the workplace and can directly click on the point and be brought to that residue. In Figure 6 we can see a textual description of the query and point in the workplace that is associated with Drosophila at index 115. The curation can then be accessed by simply right clicking the mouse and opening the curation (see Figure 5).

6 Evaluation

The CATPA method was evaluated by both quantitative and qualitative methods to determine its usability and feasibility for use as a standard protein manipulation tool. Because of space constraints, only a brief discussion will be given. A fuller discussion can be found in the full technical report. We identified two primary dependent variables:

T1 Efficiency - the time it takes a user to solve a particular problem.

T2 Accuracy - the degree of correctness that users have to solve the problems.

	Male/Female	Age ≤25/>25	Computing expert/novice	Biology expert/novice
CATPA	18/8	22/4	10/16	24/2
PFAAT	17/4	17/4	9/12	20/1

Table 1: Demographic composition of subjects

Efficiency was measured using the time taken to write each query. To measure accuracy, we adapted the procedure and coding scheme from [13]. For each query we identified errors in the following categories: (i) terminology comprehension errors (ii) positioning errors, (iii) identification errors and (iv) interface usage errors.

A number of independent variables (factors) could be used to study effect on efficiency and accuracy. For our study, the primary factor was the type of software. All of the possible factors that could be considered are as follows: **F1** Type of software - PFAAT or CATPA; **F2** Type of question - browse, search or annotation/curation; **F3** Experience of subjects (novice, or experienced); **F4** Sex of subjects (male or female); **F5** Age of subjects.

The two primary research questions we wanted to answer were:**R1** Does the efficiency of the user depend on the use of one of the two software systems? **R2** Does the accuracy of the user depend on the use of one of the two software systems? Based on the other factors, we can also answer interesting other questions: **R3** Does the efficiency of the user vary in either or both softwares based on the type of task? **R4** Does the accuracy of the user vary in either or both softwares based on the type of task? **R5** Does the efficiency or accuracy of the users vary in either or both softwares based on any other factor (such as sex, age, experience)?

We conducted an experiment (IRB Study# 03-8556) to answer these research questions. Subjects were required to formulate queries in one of the two software systems described above. The type of software (PFAAT vs. CATPA) was used as a between-subjects factor, while the type of task (search, browse, annotate) was a within-subjects factor [15, Chapter 5].

6.1 Subject Composition

Subjects were recruited from undergraduate students in the school of Informatics. A total of 47 subjects participated in the study. They were randomly assigned one of the two software systems. After the assignment, 26 subjects completed the tasks for CATPA, and 21 subjects completed the tasks for PFAAT. The students were given refreshments and a nominal payment for participating in the study. All the subjects had taken basic computing courses, including some programming courses. Table 1 shows some demographic composition of the subjects.

From the above table, the subjects were uniformly divided between computing experts and novices. The composition of age, sex and biology experience was typical at this level in the university. Hence, factors F3 through F5 were not considered for further analysis.

6.2 Experimental Tasks

A set of seven tasks were created at various levels of complexity for the subjects to perform. The tasks were of three major categories: (i) basic I/O (input and output), (ii) annotation or curation, and (iii) query. Such tasks form the core of what biologists typically are interested in performing using biological software. The reasoning behind choosing these tasks were three-fold. First, biologists need to either manually transfer data to Internet-based applications, or installing data-intensive applications locally to perform their analysis, so I/O is indeed a major activity performed by the scientists. Secondly, we are interested in the process of

Task no.	Description	Category	Complexity
1	Aligned protein data import	input/output	low
2	Saving alignment and annotation updates	input/output	low
3	Annotation creation	annotation/curation	medium
4	Loading annotated alignment data	input/output	low
5	Altering position of annotation information	annotation/curation	complex
6	Annotation search with result hint	query	medium
7	Residue search without result hint	query	complex

Table 2: Usability evaluation tasks - in the order presented to the participants

curation which is a significant task solved by CATPA. Third, the way scientists perform their work is by performing searches on similar experimental results, and hence querying forms a core functionality that need to be tested. Users were given paper copies of the tasks, and the timing system (described below) also shows the tasks on the screen. The wordings of the tasks were properly adapted to the corresponding software system. The tasks were presented in sequence since they had a certain degree of continuity between them. The tasks involved working with a real-life protein family data including annotations and curations associated with specific residues. Table 2 shows the sequence, type and complexity of the tasks.

6.3 Experiment Procedure

The experiment was performed in a computing lab. Both the groups were in one room where they were given a short discussion on the study and its purpose. The subjects were then divided into two groups where they were given a 15 minute introduction to the software systems and shown some tasks similar to the ones in the actual experiment. They were then given access to one of the two systems chosen at random. CATPA users had direct access to the central CATPA database, while the PFAAT users were given access to files that they could import into the software. Although the formats were different, the groups had access to the same genetic data, including the same sequences and alignment data.

The main body of the experiment, which lasted about one hour, involved subjects performing a set of seven tasks using the software interface that they received. Each subject had to work with two software windows: the primary window where they used the tested software interface, and another window where they were allowed to enter the results they obtained from the task (or indicate that they finished the task) and advance to the next task. This second window was driven by a timing application, which kept track of the answers given by the subjects, as well as the time taken by the subjects to answer each task. To avoid errors in understanding the timing interface, a dummy task was added at the beginning which allowed subjects to get familiar with the timing interface before starting the real tasks.

At the conclusion of the actual experiment, the tasks were coded for efficiency evaluation. Each response was assigned a value between 0 (indicating completely erroneous) to 5 (no errors). A coding scale was created based on the types of errors discussed earlier. To avoid bias in the efficiency measures, two independent evaluators were used to evaluate the tasks. The inter-evaluator correlation coefficient was computed to be 0.87, which was an acceptable high value. The average of the two evaluation values were used for the analysis.

A. Results on efficiency (in seconds) for each task. Lower mean values indicate more efficient.

Task no.	CATPA mean	PFAAT mean	ANOVA sig.	M-W sig	Significance
1	190.9	331.1	0.074	0.000	**significant**
2	136.5	104.8	0.244	0.120	not significant
3	300.9	176.8	0.031	0.059	not significant
4	74.0	51.0	0.171	0.275	not significant
5	61.2	148.8	0.030	0.003	**significant**
6	140.2	404.7	0.000	0.000	**significant**
7	57.4	83.9	0.033	0.003	**significant**
avg.	137.3	185.9	0.035	0.000	**significant**

B. Results on Accuracy for each task - higher mean values indicate better accuracy.

Task no.	CATPA mean	PFAAT mean	ANOVA sig.	M-W sig	Significance
1	4.19	3.67	0.08	0.100	not significant
2	3.88	3.71	0.313	0.308	not significant
3	3.92	3.52	0.097	0.132	not significant
4	3.88	3.20	0.030	0.038	**significant**
5	0	0	-	-	-
6	3.35	3.00	0.233	0.323	not significant
7	3.31	3.14	0.388	0.484	not significant
avg.	3.21	2.89	0.138	0.274	not significant

Table 3: Significance results obtained from the empirical evaluation

6.4 Results

We first analyzed the data using one-way ANOVA for each of the seven different tasks as well as on aggregate. However, normalcy tests revealed that the data violated some of the assumptions necessary for the use of parametric tests. Hence we also evaluated the data using the conservative, non-parametric Mann-Whitney test [29]. 1-tailed significance values for both tests are reported here, although we use the Mann-Whitney test findings to determine significance. Table 3 shows the findings of the analysis. The efficiency (timing) data is reported in seconds. Lower timing data indicates higher efficiency, and higher accuracy values indicate better performance.

The Accuracy measures of task 5 could not be taken because of an experimental error. Among the rest, CATPA was significantly better than PFAAT in efficiency for four tasks, in accuracy for one task, and in overall efficiency. In only one task (task 3) PFAAT was better in efficiency, although margninally non-signifcant at 0.05 level of significance.

The analysis results suggest some highly interesting conclusions:

1. Complex querying tasks were significantly quicker to perform using CATPA, because of its improved and more powerful interface.

2. Saving alignments and reloading them may cause mental conflicts in users' minds about where the files are actually saved, whereas the database save feature in CATPA reduces such conflicts, which results in more efficiency from users.

3. Interestingly, a user interface capturing much more functionalty does not significantly affect the accuracy of the users. Users were equally as accurate in both the systems.

7 Summary and Conclusion

We have reviewed bioinformatics systems that provide excellent mechanisms for annotation, but have observed that the experimental biologist has been left behind in the race to understand the ever increasing number of sequences. From this observation, we were motivated to design a bioinformatics system for experimental biologists that provided the same kinds of functionality found in their annotative counterparts. This system, CATPA, seems to provide most of what experimental biologists need. To check whether our conjecture was true, we performed a usability evaluation on CATPA by comparing it with PFAAT, another tool for protein annotation. We performed both quantitative and qualitative evaluations. The quantitative analysis showed CATPA to be significantly faster than PFAAT for curation editing and searching operations. The qualitative analysis demonstrated that CATPA did provide appropriate visual cues to the users for improving speed and accuracy of operations. The verbal protocol data obtained from the qualitative analysis indicated areas of improvement in CATPA. We believe that CATPA has potential to make an important impact on experimental biologists. We will focus next on adding other functionality including the ability for systems to query each other peer-to-peer.

References

[1] T. Bailey and C. Elkan. Fitting a mixture model of expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, 1998.

[2] L. Bainbridge. Verbal protocol analysis. In J. Wilson and E. Corlett, editors, *Evaluation of Human Work, A practical ergonomics methodology*, pages 161–179. Taylor and Francis, 1990.

[3] A. Bateman et al. The pfam protein families database. *Nucleic Acids Research*, (302):276–280, 2002.

[4] A. Baxevanis. The molecular biology database collection: 2003 update. *Nucleic Acids Research*, 31(1):1–12, 2003.

[5] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, and "et.al". Genbank. *Nucleic Acids Research*, (30):17–20, 2002.

[6] E. Birney, M. Clamp, and T. Hubbard. Databases and tools for browsing genomes. *Annu. Rev. Genomics Hum. Genet.*, (3):2293–310, 2002.

[7] B. Boeckmann, A. Bairoch, R. Apweiler, M. Blatter, A. Estreicher, E. Gasteiger, E. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Research*, (31):365–370, 2003.

[8] P. Booth. *An Introduction to Human-computer Interaction*. Laurence ErlBaum Associates Publishers, 1989.

[9] P. Bork and E. Koonin. Predicting functions from protein sequences–where are the bottlenecks? *Nat. Genet.*, (18):313–318, 1998.

[10] S. Brenner. Errors in genome annotation. *Trends Genet.*, (15):132–133, 1999.

[11] L. L. Conte, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2002: refinements accomodate structural genomics. *Nucleic Acids Research*, (30):264–267, 2002.

[12] M. O. Dayhoff, editor. *Atlas of Protein Sequence and Structure*, volume 5. National Biomedical Research Foundation, Washington, DC, 1972.

[13] P. De, A. Sinha, and I. Vessey. An empirical investigation of factors influencing object-oriented database querying. *Information Technology and Management*, (2):71–93, 2001.

[14] D. Devos and A. Valencia. Intrinsic errors in genome annotation. *Trends Genen.*, (17):429–431, 2001.

[15] R. E. Eberts. *User Interface Design*. Prentice Hall, 1994.

[16] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. Sigrist, et al. The prosite database, its status in 2002. *Nucleic Acids Research*, (30):235–238, 2002.

[17] L. P. Freedman, editor. *Molecular Biology of Steroid and Nuclear Hormone Receptors*. Birkhäuser, 1998.

[18] Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.*, (11):1425–1433, 2001.

[19] D. Higgins and W. Taylor, editors. *Bioinformatics sequence, structure and databanks*. Oxford University Press, Oxford, UK, 2000.

[20] L. Iyer et al. Quod erat demonstradum? the mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Bio.*, 2(12):1–11, 2001.

[21] J. J. Johnson, K. Mason, C. Moallemi, H. Xi, S. Somaroo, and E. Huang. Protein family alignment annotation tool. *Bioinformatics*, 19(4):544–545, 2003.

[22] I. Jonassen, J. Collins, and D. Higgins. Finding flexible patterns in unaligned protein sequences. *Protein Science*, (4):1587–1595, 1995.

[23] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The kegg database at genomenet. *Nucleic Acids Research*, (30):42–46, 2002.

[24] V. Kunin, I. Cases, A. J. Enright, V. de Lorenzo, and C. A. Ouzounis. Myriads of protein families, and still counting. *Genome Biology*, 4(401), 2003.

[25] D. W. Mount. *Bioinformatics sequence and genome analysis*, chapter 3. Cold Spring Harbor Laboratory Press, 2001.

[26] D. Norman. *The Design of Everyday things*. Doubleday Currency, 1990.

[27] L. Rigoustos and A. Floratos. Combinatorial pattern discovery in biological sequences: The teiresias algorithm. *Bioinformatics*, (14):55–67, 1998.

[28] S. Salzberg, D. Searls, and S. Kasif, editors. *Computational Methods in Molecular Biology*. Elsevier, 1999.

[29] S. Siegel. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, 1956.

[30] C. Sigrist, L. Cerutti, N. Hulo, et al. Prosite: a documented databases using patterns and profiles as motif descriptors. *Brief Bioinformatics*, (3):265–274, 2002.

[31] G. Stoesser, W. Baker, A. "van den Broek", E. Camon, M. Garcia-Pastor, et al. The EMBL nucleotide sequence database. *Nucleic Acids Research*, (30):21–26, 2002.

[32] The FlyBase Consortium. The flybase database of the drosophila genome projects and community literature. *Nucleic Acids Research*, (31), 2003. http://flybase.org.

[33] G. Thijs, M. Lescot, K. Marchal, S. Rombaut, et al. A higer order background model improves the detection of regulatory elements by gibbs sampling. *Bioinformatics*, 17(12):71–93, 2001.

[34] J. Westbrook et al. The protein data bank: unifying the archive. *Nucleic Acids Research*, (30):245–248, 2002.

[35] C. Wu et al. The protein information resource. *Nucleic Acids Research*, (31):in press, 2003.

[36] C. H. Wu, H. Huang, L.-S. L. Yu, and W. C. Barker. Protein family classification and functional annotation. *Computational Biology and Chemistry*, (27):37–47, 2003.