

Sticks, balls or a ribbon? Results of a formative user study with bioinformaticians

Olga Kulyk¹, Ingo Wassink¹, Paul van der Vet¹, Gerrit van der Veer^{1,2}, Betsy van Dijk¹

1: Human Media Interaction Group
University of Twente
P.O. Box 217, Enschede, The Netherlands
{o.kulyk, i.wassink, p.e.vandervet, bvdijk}@ewi.utwente.nl

2: Human-Computer Interaction
Open University,
Valkenburgerweg 177, 6419 AT
Heerlen, The Netherlands
gerrit@acm.org

Abstract. User interfaces in modern bioinformatics tools are designed for experts. They are too complicated for novice users such as bench biologists. This report presents the full results of a formative user study as part of a domain and requirements analysis to enhance user interfaces and collaborative environments for multidisciplinary teamwork. Contextual field observations, questionnaires and interviews with bioinformatics researchers of different levels of expertise and various backgrounds were performed in order to gain insight into their needs and working practices. The analysed results are presented as a user profile description and user requirements for designing user interfaces that support the collaboration of multidisciplinary research teams in scientific collaborative environments. Although the number of participants limits the generalisability of the findings, the combination of recurrent observations with other user analysis techniques in real-life settings makes the contribution of this user study novel.

Keywords: User study, bioinformatics, human-computer interaction, co-located collaboration, visualisations

1 Introduction

The importance of bioinformatics in life science domain has grown tremendously over the past years and is expected to do so for years to come. Various experts have to collaborate and to work with shared knowledge. They are forced to use complex scientific applications that require expertise they often do not have. Currently used bioinformatics interfaces are designed for expert bioinformaticians, cheminformaticians and computational biologists. They are too complicated for novice users such as bench biologists [1]. The users' cognitive load is overstretched by huge amounts of heterogeneous data, mutually inconsistent representations, and the complexity of and limited interaction with the user interfaces of bioinformatics tools. A new generation of interactive visualisation interfaces has to meet user requirements as well as to improve the exploration of large amounts of heterogeneous data and to enhance knowledge construction [2]. Therefore, there is a need for user-centred interface design and evaluation in order to improve the effectiveness and efficiency of both visualisation systems and bioinformatics tools.

Understanding the users in their context of work, how and why they approach use different information resources and tools, is essential to provide information technology, in particular interactive visualisations [3]. In the life science domain, interactive visualisations are used to facilitate data analysis and hypothesis formation. A user interfaces and visualisation project at

the University of Twente within the BioRange project¹ is devoted to the user-centred design and evaluation of visualisations and enriched interactions in order to enhance the exploration of bioinformatics resources by multidisciplinary teams of scientists. User studies will help to overcome the barrier between non-experts and the available bioinformatics resources, and therefore will enhance the knowledge discovery process.

The purpose of this empirical user study is to explore working practices and experiences of users from different bioinformatics sub-domains and disciplines with various levels of expertise in real-life settings. We also aimed to identify the key aspects and user requirements in the context of scientific collaborative environments. Such an environment contains high-tech devices, such as large displays for interactive visualisations and digital whiteboards. Therefore, we started to analyse the current working style of multidisciplinary project teams in real-life contexts in order to understand the target user group.

The remainder of the report is organized as follows. The next section contains a brief review of related formative user studies in bioinformatics. The third section describes our method and three main target groups. Then, the results presented as user profile descriptions and design implications are discussed, followed by conclusion and discussion.

2 User and task analysis in bioinformatics

Most published studies focus on the evaluation of the existing tools (e.g., [4]) but not on user analysis to formulate requirements. There are very few user analysis studies in the life science domain available in the literature, and none concentrating on multidisciplinary collaboration in bioinformatics. Dunbar performed ethnographic observations and interviews to study cognitive mechanisms and complex thinking, albeit in molecular biology [5]. The only user analysis study in the bioinformatics domain we are currently aware of is the study reported by Barlett and Toms [6]. They proposed an information behaviour framework integrated with task analysis for studying patterns among bioinformatics experts. Their work is based on 20 interviews with bioinformatics analysts working on functional analysis of a gene [6].

Previous studies on creative and complex thinking of life scientists have shown that multidisciplinary in research teams stimulates the process of creative thinking and reasoning [5, 7]. Creativity may be stimulated by providing an interactive environment and an appropriate context to scientists [9]. According to creative thinking theory, there are three stages of creative problem solving: preparation, production and judgement [10]. Visualisations and tentative interactions can support creativity in all three stages. However, they are especially important in the production stage to support the generation of multiple hypotheses. The challenge at the judgement stage is to design visualization for an optimal perception of the information [10]. User interfaces and visualisations are part of the problem solving process. We will need to test and optimise the visualisation designs and interaction styles by performing user analysis and iterative evaluations [3, 11]. Collaborative creativity involves both individual and group working practices, which introduces a new level of complexity in understanding the target users and designing for their needs [8].

¹ BioRange is a research activity of the Netherlands Bioinformatics Centre (NBIC). BioRange aims to enhance collaboration between bioinformatics and genomics research and to stimulate bioinformatics development in the Netherlands.

3 Method

We conducted user analysis studies in bioinformatics in order to gain insight into the needs and working practices of researchers from different sub-domains. These studies included a questionnaire, ethnographic observations and interviews. Different target groups were chosen for the study in order to get different perspectives on users in the bioinformatics domain. For each user group, a different method of study was chosen, based on both the goal of the analysis and the characteristics of the target users.

3.1 Questionnaire

The first target group consisted of novice users. The aim of this part of our study was to gain more insight into how these users deal with bioinformatics problems and how do they use bioinformatics resources: What is their working strategy? What is their strategy of getting from the target question towards a conclusion? If they draw conclusions, do they use additional information to verify them? A multidisciplinary group of students taking a nine weeks introductory bioinformatics course at the Bachelor's level offered by the CMBI, Radboud University, Nijmegen, the Netherlands, participated in the questionnaire part of this user study. They had no experience with bioinformatics tools and therefore had no formed opinion about the usefulness of bioinformatics interfaces. A discussion of the results can be found in section 4.1. The questionnaire and its results are presented in Appendix I.

Prior to the questionnaire, regular contextual, unobtrusive observations were performed during a weekly bioinformatics course. The environment where students had weekly practical course consisted of the multiple rows of tables with PCs (see Figure 1). During this course, students learned how to use different types of bioinformatics resources. First we wanted to gain insight into the daily practice of the novice users while they learn to use different bioinformatics tools and deal with the real-life problems.



Figure 1: Observations of the novice users took place in the practical bioinformatics course room at the Radboud University of Nijmegen

The collected observations were translated into simple statements about the way in which novice users deal with practical bioinformatics problems using different on-line web resources, both data and tools. Based on these statements, a questionnaire was designed to check and refine the statements. In order to correlate the questions with students' recent practice, 3D visualisations of a familiar protein were included (Figure 4.A). Students had to apply knowledge from the whole course and use different bioinformatics databanks and tools for this assignment.

3.1.1 Participants

In total 47 (21 female and 26 male) students took part in the user study. The participants were mainly Dutch and German students of the Radboud University of Nijmegen. The students had different backgrounds (molecular science, chemistry and general natural science) The average age of participants was 21.5 years. Based on the user profile questions it became clear that students' level of experience with software tools is generally quite high. The majority of students used the Windows platform and multiple mail programs, web browsers, search engines, text editors, spreadsheets and instant messengers.

3.1.2 Procedure

A pilot test with two course assistants was conducted. Based on their feedback, the necessary adjustments to the questions and the layout were made.

Students were asked to fill the questionnaire at the end of the course day. A course assistant explained the purpose of the study and emphasised that participation is anonymous and voluntary. Students were given an introduction on how the questionnaire is constructed. It took 15-20 minutes for the students to answer the questions. The questionnaire consisted of three parts: 1) background information and general software usage questions; 2) questions on 3D visualization tools; 3) questions on the web-based databanks to obtain protein sequences data. Additional space was left for extra comments after second and third part. Fourteen out of twenty-one questions used a 5-point Likert-scale, where '1' was presented as 'Agree strongly' and '5' as 'Disagree strongly'. Three questions were single-choice questions and another two were multiple-choice questions. The last two questions were ranking questions, where users had to rank the options by importance on a scale from 1 to 3. The response of the questionnaire was about 90%. The full questionnaire can be found in Appendix I. The results are discussed in Section 4.1.

3.2 Ethnographic observation

The second target group consisted of multidisciplinary teams collaborating on a joint scientific experiment. Such a team consists of scientists from different domains related to bioinformatics, for example, molecular biology, chemistry and statistics. Teams of scientists are of special interest for our further studies, since, as mentioned above, creativity of scientific thinking occurs in groups rather than individuals. The goal of the observation was to gain more insight into how researchers from different disciplines collaborate while solving biological problems and how they use technologies supported by the meeting room environment.

Novice users have little or no experience in collaboration with other researchers. Therefore, a multidisciplinary team of experts was chosen for an ethnographic observation. The multidisciplinary team, consisting of three biologists, two statisticians and two bioinformaticians from different research institutes, had a regular project meeting at the University of Amsterdam, the Netherlands, in which they were discussing the p53 protein. Some participants were direct colleagues of each other. All participants knew each other from earlier meetings of the same project. The regular meeting of a project team was audio recorded with participants' permission.

Figure 2 shows the layout of the meeting room. This meeting room was equipped with:

- A large table, with seats around it
- A video projector, remote control with a laser pointer

- A paperboard
- A wooden stick used as pointer
- A whiteboard

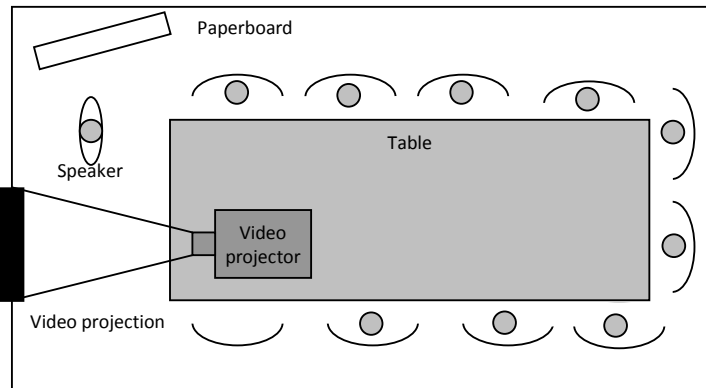


Figure 2: Layout of the room

The subject was the interpretation of the results of a microarray experiment. Details and findings are presented in section 4.2.

3.3 Interview

The third target group were bioinformaticians and were expert users of bioinformatics applications. The participants of this study were researchers with different backgrounds working in bioinformatics. Bioinformaticians were selected in order to gain insight into their experiences in collaboration with researchers from other areas and their use of different tools.

Interview questions were focused on their work practices and on the use of bioinformatics tools. Since observation showed that collaboration is essential in bioinformatics research. We also asked about experts' collaboration experience and opinion on how future technology in collaborative environments, such as large wall displays, might influence collaboration.

Semi-structured interviews were held at the the Centre for Molecular and Biomolecular Informatics (CMBI), Radboud, University Nijmegen, The Netherlands employing contextual inquiry technique [12]. In this research group, scientists with biology, molecular biology, bioinformatics and statistics backgrounds work together on various projects. So far, three researchers (two PhD students and one post-doc) were interviewed. The three participants were male and aged between 25 and 30 years, and were active in the bioinformatics domain for at least 2.5 years. The sessions were audio recorded with participants' permission. The full transcripts of interviews with these researchers are omitted here for privacy reasons. Details and findings are presented in section 4.3.

4 Results

4.1 Contextual observations with novice users and questionnaire

The unobtrusive observations of students during the practical assignments of the introductory bioinformatics course showed that students often worked in groups of two to four. on the assignments The course assistants were often asked for explanations about both the material and how to use different bioinformatics tools. The atmosphere during the classes was very

informal and active discussions were going on all the time. Students used a wide variety of different software tools simultaneously, e.g. mail program, spreadsheet, web search, messaging, games etc. In addition the electronic course material together with a paper study guide was used for practical assignments.

The exploration of 3D structures of proteins is very important in the course and was performed a lot. The students easily recognised the structure of a protein. Different tools were used for exploring the 3D structure of a protein, including Jmol (jmol.sourceforge.net) and Chime (www.mdl.com/products/framework/chime/), but Yasara (www.yasara.org) was the most popular. The right-click menu in the 3D visualization tool Yasara was not self-explaining due to the complex structure. Students are often overloaded with information, as the option to make information visible on demand is missing. They had problems choosing and switching between different views inside one tool. Users did not exactly know what to do; they were often searching for the different representations to find the information.

The interaction with 3D structures of proteins was limited to the keyboard and mouse. The most common interaction styles were:

- rotate
- point at node (e.g. pointing at a protein structure to see which residue is it)
- zoom in/out
- selecting amino acids
- hiding irrelevant and showing relevant parts of a protein

Selecting amino acids was difficult to perform, because the current selection feedback was missing. The students used alternative colour-codings provided by the visualisation tools to emphasize certain amino acids.

The preferred 3D view for a complete protein is 'Ribbon', and for the part of a protein, 'Sticks' (see Figure 3). The 3D protein structure gives users the necessarily information about the function of the residue..

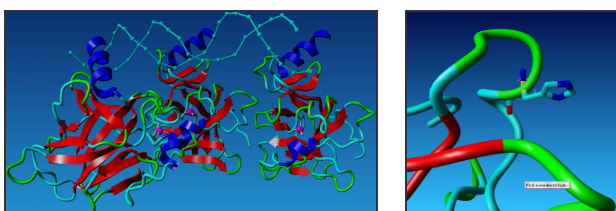


Figure 3: View Preference, 'Ribbon' (left), 'Sticks' (right)

In the use of online databases, it is also often unclear what type of search the databases support. Many non-ordered options are presented to the user to optimize the search, but also a lot of options are hidden. The absence of a history function is problematic when the users are redirected to a different web application.

While using web portals, in particular MRS (mrs.cmbi.ru.nl) and SRS (e.g., srs.ebi.ac.uk), to extract data from different databases, novice users do not change options to optimize their search. Cross references are frequently used to obtain more information. Finding reliable information is the most important criterion for users when they search for information. One-way Anova analysis (comparing means) did not show any significant differences between genders, and no significant differences between students' groups with different study

backgrounds. The complete results of the questionnaire analysis are presented in Appendix II: Table 1.

4.1.1 Usability problems

In general, users report that Yasara is currently the most useful visualization tool for a protein, and MRS is considered to be the most useful web-based search platform. After analyzing the users' additional comments and observations results, the following common usability problems were discovered:

Visualization tools: Yasara, JMol, Chime

1. Selection and visibility problems:

- a. Hard to make certain residues or a part of a protein visible. Suggestion from users: Decrease the brightness at the back of a protein in order to make it easier to distinguish the front and the back side.
- b. An option to make names of the residues visible in part of a protein is missing.
- c. Hard to make the hydrogen-bonds visible, specifically in a part between DNA and a protein.

2. Linking between views problem:

- a. No linking between overview and detailed view. Suggestion from users: include transparency option in order to see more layers of the protein simultaneously.

3. Interaction problems:

- a. 'Shift' interaction option with a 3D view of a protein is missing.
- b. Transformation, scaling and moving is difficult.

4. Learnability problem:

- a. Many options are not clear. It is not intuitive what the possibilities with the visualization tool are and how to use them. Help is missing.

Protein Sequence Databases: SRS & MRS

1. Layout problem:

- a. 'Submit' button should be on the right below, not on the navigation bar and not in the header.

2. Consistency problem:

- a. Fasta format for the protein sequence is not used in every database. This causes waste of time.

The common general remark about the course was that students were not motivated enough to explore the bioinformatics tools more than suggested in the study guide.

The obtained results from this user study provides a better understanding about the novice users' daily working practices with different bioinformatics tools. The participants of this study are quite skilled in using general software tools. However, inconsistency in representation, complexity and limited interaction with user interfaces of bioinformatics tools combined to cause information overload and time loss for users. Therefore, there is a need in user-centred interface design and evaluation in order to reduce the workload and improve the effectiveness and efficiency of both software tools and web resources use. In addition, it seems to be important to support researchers to keep track of their thoughts and ideas during the information search and analysis.

4.2 Observing a multidisciplinary life science team

The results of this field observations are based on the analysis of audio recordings and the observations. The working atmosphere was informal. During the session people were drinking coffee and having candies.

4.2.1 Project team

Within the p53 protein project, this team was testing the influence of down-regulating and up-regulating the genes possibly related to this protein. The group did the experiment without

a clear hypothesis, which is better characterised as a “try and see what happens” experiment. The type of experiments in which the hypothesis is formulated after the data are collected and analysed, is quite common in molecular biology.

The main goal at the observed meeting was to discuss the statistical steps for analyzing the obtained microarray data. The schedule of the meeting wasn't written down on a paper or projected on the screen; everybody knew a schedule implicitly. The meeting had a clearly discernible structure:

- Presenting the theory (1 hour)
- Presenting the practice (1 hour)
- Discussion (1 hour)
- What to do next (0.5 hour)

The two statisticians presented the theory and the practice of the statistical analysis. These two project members had done the practical part together as well. They used the video projector and a PowerPoint slideshow. First, one of the statisticians (statistician A) presented the theoretical part, after this, the other (statistician B) presented the practical part. During the presentation, the presenters could be interrupted for asking questions. This was done frequently. During the practical part, statistician A interrupted the statistician B many times to add additional theoretical information. One of the biologists arrived in the middle of the presentation part. The video projector was not used during the discussion.

After this, a discussion evolved about the task distribution and about the not unrelated issue of whether or not the goal of the experiment was to fit the results in a model. From the biologists' point of view, this was not the case, which was a quite normal reaction, since creating a model of the results is something biologists not always aim at.

In general, the discussion was very active. However, not all team members were always active during the meeting. It seems that this depended on the experience with the subject at hand the group member ascribed himself or herself. The biologists seemed to be the most active members. There was no appointed leader or chairman of the meeting, neither was there an appointed reporter. Everyone was responsible for making his/her own notes. The project team, and one biologist in particular, generated a lot of new ideas, but these ideas had to be worked out in subsequent research and further meetings. It was clear from the observations that the project team could not remember what was agreed on during previous meeting.

After the meeting, a smaller group moved to the office where statisticians demonstrated the new possibilities for visualization and demonstration of the data to one of the biologists. The tools Spotfire and Gene Ontology (used for visualization of Biological Pathways) were used for that. It was observed that the statistician wanted to drag and drop different visualizations from one window to another, but the necessary linking between different applications and views was missing.

4.2.2 Using visualizations

The meeting was mainly about interpreting statistical results. The used visualizations consisted of diagrams (frequency diagrams, scatter plots), but also custom-made sketches on the paper-board for showing abstract ideas or for explaining something. To compare and to interpret diagrams, multiple diagrams were shown at the same time on a single slide. The diagrams differed in populations or in parameter settings. This made it easy to compare the results. Everyone present knew how to interpret the diagrams, but the statisticians were the

only ones who knew how to create the diagrams using statistical models such as Anova. A visualisation based on a clustering technique was also used. However, everyone agreed that this kind of visualization was nice to see, but nobody knew how to interpret these visualisations and how to assess whether they were useful or not.

The use of technology supported by the environment:

Video projector/PowerPoint

- The video projector was used for giving the presentation about the theory and for showing the results of the practice. The pictures shown were static images. PowerPoint was used for zooming in into the pictures, although this tool does not suit this kind of interaction.
- The slideshow was made available for the participants after the meeting.

Paperboard

- The paperboard was intensely used during the presentation and the discussion. Meeting participants used the paperboard as a big notepad. They looked back to previous pages and they used new blank papers for overwriting/clean up certain parts of used papers.
- The paperboard was also used for writing down the schedule, task distribution etc. These papers were kept after the meeting, but the notes on it were not mailed to the other participants.
- Sketches that statisticians drew using the paperboard where schematic-based instead of text-based.

Whiteboard

- There was a whiteboard in the room, although it was not positioned on the wall. The whiteboard could be placed in the area where the video projector, what resulted in using either the video projected or the whiteboard. So in this case, the whiteboard was not used.

Other tools

- The windows calculator projected by the video projector was used to perform calculations.
- No tools for doing experiments were used during the meeting.

4.3 Interviews

Bioinformaticians live between two worlds biology and computer science. They have the necessary knowledge to collaborate with biologists, which is something computer scientists cannot do or at least not so smoothly. Bioinformaticians develop and use tools to collect huge amounts of data from (online) databases and to analyse these data using statistical techniques. As one respondent said, “bioinformaticians are not computer scientists, who can build large software architectures, but they know how to program tools to extract biological meaning from databases”.

4.3.1 Multidisciplinary research

The researchers are very often working in multidisciplinary teams. These teams consist of researchers with backgrounds in biology, bioinformatics, chemistry, mathematics or statistics. They are also often collaborating with industries that are highly interested in this kind of

research, such as pharma and food industries. These industries pose abstract research questions, which are translated by a team leader into several concrete research questions. The interviewees highly value the collaboration in the same working space, but reject the idea of distance collaboration based on virtual meetings.

4.3.2 Tools used by bioinformaticians

Statistical analysis is vital in bioinformatics research. Huge amounts of data stored in databases are compared using statistical software. Based on these studies, conclusions are drawn. One respondent said that at his department, the staff mainly uses MatLab (www.mathworks.com/products/matlab) for doing statistical analysis. The R package (www.r-project.org) is a favourite statistical tool for many other bioinformaticians. In addition, bioinformaticians use tools designed specifically for biological data analysis. For example, tools are used for finding proteins with similar sequences or for visualizing protein structures.

The tools often provide a lot of parameters to customise their working. Although these parameters make the tools flexible, they also increase the complexity of the tools. A respondent said that most tools are very complex due to the number of parameters that can be changed. Often, this is unavoidable. Only good documentation can help in understanding of the tool, but most tools lack this. Another respondent remarked that biologists use the default parameter settings most of the time, because they do not have much knowledge about the meaning of the parameters. Bioinformaticians have a different work style. First, they try things out to verify a hypothesis or hunch, using the default parameters. If the hypothesis is more or less confirmed, then they fine-tune the parameters to optimize the results.

5 User profiles

The user profiles represent two types of researchers using bioinformatics tools: biologists and bioinformaticians. Bioinformaticians can be seen as *domain experts* in this case: they know their way around in the vast (and growing) space of online bioinformatics resources, and they know about data handling and the operation of bioinformatics databases [13]. Unlike bioinformaticians, biologists are mostly *novice users* of bioinformatics resources for their research. They are experts in doing wet-lab experiments.

5.1 Novice users

Novice users of bioinformatics tools, such as biologists lack the programming skills that expert bioinformaticians have. They often do not directly understand how programs work. As a result, they are often discouraged from experimenting with these tools. One of the interviewees stated that biologists for this reason use only default parameters most of the time. The questionnaire results show that less than 22% of novice users change parameters to assess parameter influence on the result of an experiment.

Novice users, however, are quite skilled and advanced in using general software tools. More than 68% of the participants often use cross references for getting more details about experimental results. Therefore, it is essential to provide an option to make information on demand visible. In addition, a bioinformatics database needs to clearly inform the users about what type of data it provides. Novice users get confused by many unstructured configuration options. They also miss a history function when they are redirected to a different web application.

5.2 Domain experts

The bioinformaticians know their way around in the vast (and growing) space of online bioinformatics resources, and they know about data handling and the operation of bioinformatics databases [13]. Domain experts use diverse databases and tools to collect and to analyse huge amounts of data and to draw conclusions from them.

Bioinformaticians have programming skills, and consequently, they understand how programs and tools work and they often know how to extend them. The interviewees, who are domain experts themselves, explained that bioinformaticians create and use bioinformatics tools to collect huge amounts of data from databases. This makes them less afraid to experiment with different tools and with parameter settings. Their work style can be roughly characterised as follows: they first try things out to verify their hypotheses using the default parameters. When the hypothesis is more or less confirmed, they fine-tune the parameters to optimise the results. Bioinformaticians prefer console applications over equivalent GUI or web-based interfaces, since a console allows them to customise all parameters. They also claim a console gives them more insight into how the tool works. To do the statistical analysis, they use software such as Matlab and R. Specialized software such as ClustalW (e.g., www.ebi.ac.uk/clustalw), WU-Blast2 (e.g., www.ebi.ac.uk/blast2) and Yasara are used for respectively, protein sequence comparison, sequence similarity search, and the visualisation of protein 3D structures.

When the interviewees were asked how they become aware of the existence of new tools, they mentioned that colleagues are important sources of information for learning how to use new bioinformatics tools. If they find a new tool by themselves, they test the tool and compare the results with those of familiar tools in order to establish a quality measure. Experience with, trust in, and perceived quality of tools are exchanged among bioinformaticians.

5.3 Multidisciplinary teams

In the biomedical domain, researchers very often work in *multidisciplinary teams*. These teams consist of researchers with backgrounds in biology, bioinformatics, (bio)chemistry, mathematics and statistics, but also industries are often involved in research projects.

6 User requirements

The results of the questionnaire, the observation and the interview provide useful initial input for designing interfaces to support co-located collaboration. The combined information from the three methods is translated into a set of requirements for visualisations, collaboration and multidisciplinary teamwork support in a scientific collaborative environment.

6.1 Visualisations

Researchers in this area use different types of visualisations for different types and different amounts of data:

- 3D visualisations of a protein are used for searching amino acids which could possibly be involved in the protein's function (see Figure 4.A). This activity requires intensive interaction with the 3D model of a protein, consisting of zooming and selecting different levels of detail for the whole 3D model or just parts of it. Therefore, *easy switching between views* and *feedback about current selection* are essential.

- Another frequently used visualisation is a sequence alignment visualisation (Figure 4.B). In alignment visualisation, the similarity of two or more proteins or DNA sequences is depicted by means of a colour-coded map that shows a number of strings (amino acids in the case of proteins, nucleotides in the case of DNA) that are aligned to achieve optimum similarity over the entire string.
- The multidisciplinary team was discussing a micro-array analysis during the regular project meeting. Micro-array analysis is a quantitative method to study the simultaneous activity of thousands of genes at a certain point in time. There are two types of micro-array studies: in one, absolute gene activity is measured for a particular cell while in the other, gene activity of cells under different conditions (for example, from sick and from healthy tissue) is compared. Because the raw data from a micro-array experiment are normally marred by a lot of noise, statistical analysis is used. Graphical plots of the results are used to aid the analysis (see Figure 4.C).

Visualisations are very important in bioinformatics. One of the interviewees mentioned that visualisations are often underestimated in biology and suggested they should also be used for showing active and inactive parts of biological networks.

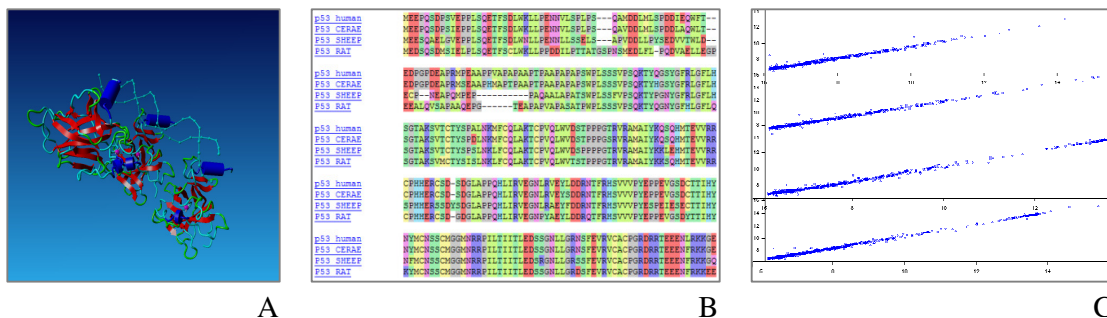


Figure 4: Visualisation used for: A) protein structure, B) sequence alignment, C) micro-array analysis

6.2 Collaboration and multidisciplinary teamwork

Collaboration can be performed in different forms, ranging from working together in the same physical place (meeting room but also at the same work floor) and at distance by publishing their work and reading those of others.

Bioinformatics researchers have to work with people from other fields of expertise, such as biology and statistics, because they do not have these skills themselves. This was not only found from the observation of a multidisciplinary team, but also confirmed by the interviewees.

6.3 Scientific collaborative environment

During project meetings, experiments are difficult to perform, because they often take too much time. Therefore, the collaboration environment will mainly be used for the *discussion of (intermediate) experiment results* and of the *project progress*. It was clear from our observations that the project team could not remember what was agreed at previous meetings. Therefore, it is important to *assist the project teams during group discussions* to keep track of their decisions, action points and ideas.

New technologies offer the opportunity to enhance meetings of the multidisciplinary teams by means of a scientific collaborative environment, such as large interactive displays [13].

These large interactive displays can be used for discussing the setup of an experiment and for sharing and joining interpretation of experimental results, among other uses. A large display can show multiple views of the same datasets or the same type of view of different datasets for comparing experimental results. The interviewees think this can enhance creativity and stimulate discussion, although such displays should not overload users with a lot of results shown at the same time.

Although researchers can access their data for presentation during the meeting, one interviewee mentioned that preparation of a meeting will still remain important. In addition, a moderator should lead the discussion to prevent the meeting from becoming chaotic.

7 Conclusion and discussion

Most bioinformatics tools are very complex, even for domain experts, due to the number of parameters that can be set and the lack of documentation to assist users in understanding the interface. Visualisation of biological data is very important in bioinformatics field. Visualisation is used for discussing the design of an experiment and/or (intermediate) results and for assessing the progress of an experiment. However, domain experts think that the use of visualisations is currently underestimated in bioinformatics.

A lot of research is done on virtual collaboration, where the scope is on distance collaboration. Multidisciplinary collaboration is an essential part of bioinformatics research. However, Bioinformatics researchers themselves are sceptical about the idea of virtual collaboration. They expect people to be more dynamic in a joined physical space, than in virtual space. Therefore, our focus is on co-located collaboration in a scientific collaborative environment. The target group for this environment will consist of multidisciplinary scientific teams. Such an environment will contain large interactive displays for presenting experimental results or project progress in order to improve collaboration. Domain experts believe that such an environment can help collaboration, although facilitating the discussion by a moderator is essential.

Although the number of participants limits the generalisability of the findings, the combination of regular observations with other user analysis techniques in real-life settings makes the contribution of this user study novel. Further studies with a larger sample from a more diverse population will reduce the current limitation. The presented user analysis approach can be used to study multidisciplinary teams in other domains. The results of our questionnaire will also be used to improve the introductory course of bioinformatics.

The requirements above have to be validated with users in conjunction with a task model representing the current work in bioinformatics. One of the intentions is to design scenario descriptions based on real-life tasks performed by a multidisciplinary team of experts assembled in a collaboration room. By observing the teams of scientists we hope to understand their working style, ways of using the technology, visualisations and interactions styles.

The questionnaire and its detailed results, including user preferences and usability problems with commonly used visualisation tools and web-based databanks, can be found in the appendices I and II.

Acknowledgments

We are indebted to all participants of this user study, the course practical assistants for participating in the pilot test, Gert Vriend, Han Rauwerda and Timo Breit for their support.

We are also grateful to our supervisors and colleagues for their feedback.

This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

Part of this research was informally distributed among the participants of the *British HCI'06 International Workshop on Combining Visualisation and Interaction to Facilitate Scientific Exploration and Discovery*, chaired by T. Adriaansen and E. Zudilova-Seinstra.

Appendix I. Questionnaire

Below follows an exact copy of the questionnaire as it was distributed among the bioinformatics students.

Please fill in the first 4 questions and then read the next block. **It is important that you answer all of them.** Thank you.

You may answer the questions in Dutch

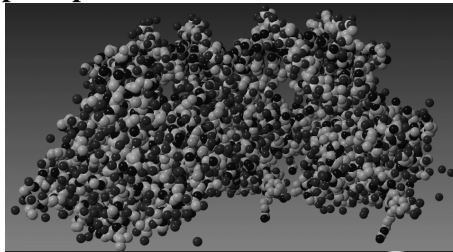
- 1 Age:.....
- 2 Gender:
 Female
 Male
- 3 Study background:.....
- 4 Which software tools do you use?
Please underline which software you use or specify which other program you use:
 - a. Operating System: Windows / Linux / Unix / Apple
 - b. Text editor: Word / OpenOffice / LaTeX / Other
 - c. Spreadsheet: Excel / OpenOffice / Lotus Notes / Other
 - d. Browser: Internet Explorer / Firefox / Netscape / Opera / Other.....
 - e. Mail program: Outlook Express / Outlook / Bat / Eudora / Thunderbird / Hotmail /GMail / Other
 - f. Search Engine: Google / Altavista / Yahoo / Other
 - g. Other frequently used software:

On the following pages you will be asked questions about your experiences regarding use of bioinformatics applications during the course. There are no right or wrong answers; we are interested in your personal opinions and experiences. Do not think about questions for a long time, but try to rely on your first reaction. It is no problem if you are not sure about this. Just try to give the answer that *you* think is most suitable. This questionnaire is completely anonymous and the results will not be associated with your name.

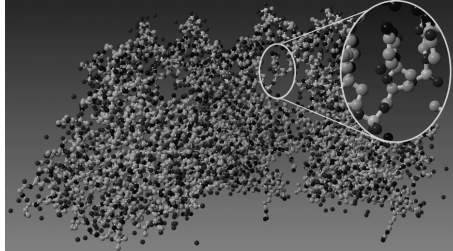
Part I: 3D Visualization tools: Yasara, JMol, Chime

1 I prefer the following 3D view of a **complete protein** structure:

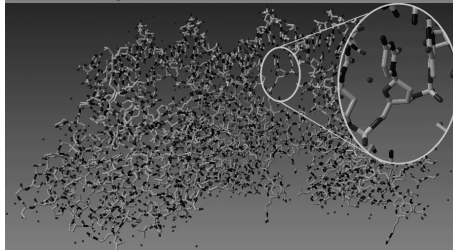
Balls



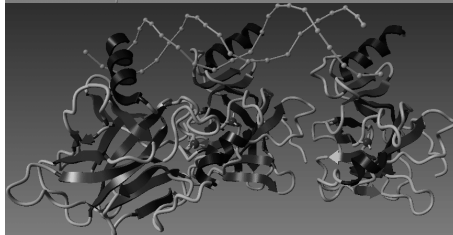
Balls and sticks (See the enlarged part -->)



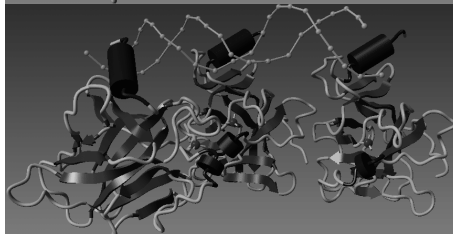
Sticks (See the enlarged part -->)



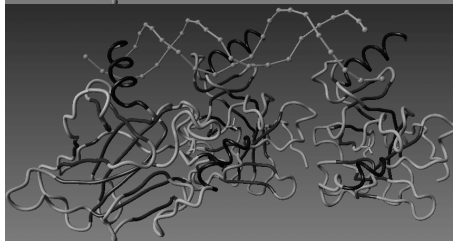
Ribbon



Cartoon



Tube



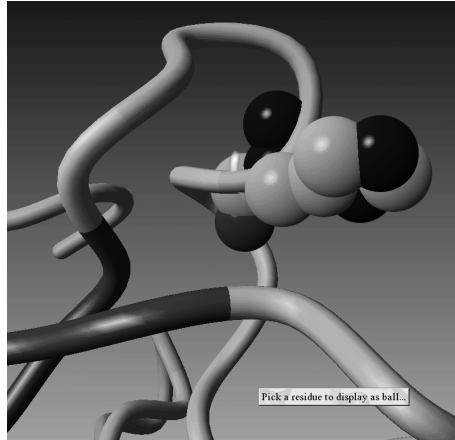
I have no preference

It depends on (please specify)

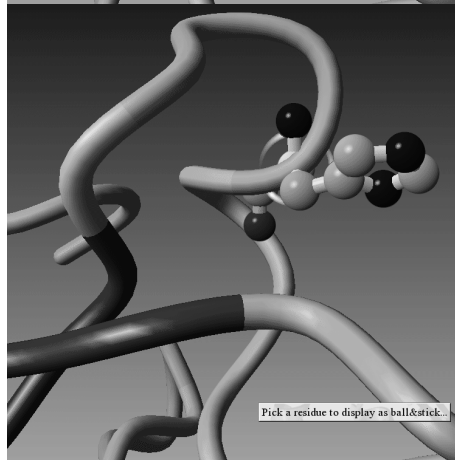
.....
.....

2 I prefer the following 3D view for a **part of a protein** structure:

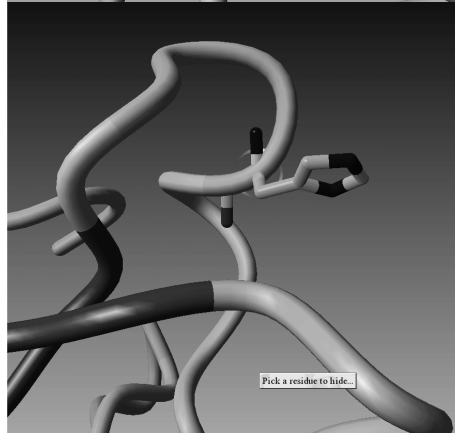
Balls



Balls and sticks



Sticks



I have no preference

It depends on (please specify)

.....
.....
.....

3 I often use an option to make only a selection of the protein structure visible.

disagree strongly disagree neutral agree agree strongly

- 4 It is easy to recognize residue's structure in a protein.
 disagree strongly disagree neutral agree agree strongly
- 5 I often use several 3D visualization tools (Yasara, Chime, JMol) to get more insight into a molecule.
 disagree strongly disagree neutral agree agree strongly
- 6 I often make only the required information of an interesting residue visible (for example, only side chain of the interesting residue).
 disagree strongly disagree neutral agree agree strongly
- 7 The 3D structure of a protein gives me information about what the function of a residue is.
 disagree strongly disagree neutral agree agree strongly
- 8 The 3D position of a residue in a protein in combination with my background knowledge about the residue always gives me enough information to determine a possible function of a residue.
 disagree strongly disagree neutral agree agree strongly
- 9 When I know the position of a residue, I often use additional resources (like access to other databases, Google, etc.) to gain more information about the protein.
 disagree strongly disagree neutral agree agree strongly
- 10 When I know the function of an **entire protein**, I often use additional resources (like access to other databases, Google, etc), to verify my conclusions.
 disagree strongly disagree neutral agree agree strongly
- 11 When I know the function of a **part of the protein**, I often use additional resources (like access to other databases, Google, etc), to verify my conclusions.
 disagree strongly disagree neutral agree agree strongly
- 12 I search first for existing information about the protein in sources on the internet before trying to discover more about the protein myself.
 disagree strongly disagree neutral agree agree strongly
- 13 I often use an option to hide some irrelevant part of the protein.
 disagree strongly disagree neutral agree agree strongly

Please write here your extra **comments** about the about the 3D Visualization tools:

Part II: SRS & MRS

1 I often change search options (e.g. “**Blast options**” for BLAST search, as on the figure below) to optimize my search.

Run Blast

Protein Query

Enter one or more protein sequences in *FastA* format

Or enter filename: Browse...

Protein Databanks

Choose the databank to search

Optionally enter an MRS query to limit the search space

Blast Options Advanced Blast Options

Filter query sequence (low complexity)

Scoring matrix

E-value cutoff

disagree strongly disagree neutral agree agree strongly

2 I use the extended query form (for example in SRS to find previous annotations of a protein in order to see the history of this protein annotation, as on the figure below).

Often

Sometimes

Never

Quick Searches Select Databanks Query Form Tools Results Projects Custom Views Information

SRS

Reset search SWISSPROT

Search Options

Combine search terms with:

Use wildcards

Get results of type:

Result Display Options

View results using: or

Sequence Format:

Show results per page

Fields you can search Your search terms Create a view

In a single field, you can separate multiple values by &, |, !

Field	Search Term	Checkbox
AllText	<input type="text"/>	<input type="checkbox"/>
ID	<input type="text"/>	<input type="checkbox"/>
Accession Number	<input type="text"/>	<input type="checkbox"/>
Primary Accession Number	<input type="text"/>	<input type="checkbox"/>
Description	<input type="text"/>	<input type="checkbox"/>
Gene Name	<input type="text"/>	<input type="checkbox"/>
Keywords	<input type="text"/>	<input type="checkbox"/>
Entry_Creation Date	select <input type="text" value="1"/> <input type="text" value="Jan"/> <input type="text" value="YYYY"/> <input type="text" value="1"/> <input type="text" value="Jan"/> <input type="text" value="YYYY"/>	<input type="checkbox"/>
LastSequenceUpdate	select <input type="text" value="1"/> <input type="text" value="Jan"/> <input type="text" value="YYYY"/> <input type="text" value="1"/> <input type="text" value="Jan"/> <input type="text" value="YYYY"/>	<input type="checkbox"/>
LastAnnotationUpdate	select <input type="text" value="1"/> <input type="text" value="Jan"/> <input type="text" value="YYYY"/> <input type="text" value="1"/> <input type="text" value="Jan"/> <input type="text" value="YYYY"/>	<input type="checkbox"/>
Organism Name	<input type="text"/>	<input type="checkbox"/>
Taxon	<input type="text"/>	<input type="checkbox"/>
NCBI_TaxId	>= <input type="text"/> <= <input type="text"/>	<input type="checkbox"/>
Organelle	<input type="text"/>	<input type="checkbox"/>
ProteinID	<input type="text"/>	<input type="checkbox"/>

3 When I receive results from a tool (e.g. for sequence alignment), I try to change parameters to see how it will influence the results.

disagree strongly disagree neutral agree agree strongly

4 The tools often give cross references to other databases with additional information. I often use these cross references to get more information.

disagree strongly disagree neutral agree agree strongly

5 When I am not satisfied with the information that I find in the SwissProt database, I use the following additional sources (**more than one option** can be chosen):

- Search engines
- Cross references
- Search manually in other databases
- Knowledge from other students
- Other
-
-
- None

6 When I search for information about a disease related to a protein, I use the following additional sources to get more information (**more than one option** can be chosen).

- Search engines
- Omim cross references
- Other cross references
- Other
-
-

None

7 When I look for information, the most important to me is:

Rank the options by importance to you (1..3)

- () Detailed answers
- () Reliable answers
- () Non-redundant answers
- () Easy-to-use answers (standard format, like Fasta used in ClustalW)
- () Well-documented answers (with respect to the traceability of their origin)

8 When I use cross-references, I use the cross reference according to:

Rank the options by importance to you (1..3)

- () The kind of information I want to get
- () The reliability of the source which is going to provide the data
- () The fact that I know whether the cross-reference has been added manually
- () The fact that I know whether the cross-reference has been added automatically (e.g. by computer systems)
- () Other.....

Comments:

Please write here your extra comments about **MRS&SRS** or any other comments:

You have finished the questionnaire. Thank you very much!

Appendix II. Questionnaire Results

Table 1: Questionnaire results

Question	Scale	Mean / %	SD
Part I. 3D Visualization Tools: Yasara, JMol, Chime			
1. 3D view preference of a complete protein	Mult. choice		
1a. Balls		0%	
1b. Balls and sticks		11%	
1c. Sticks		26%	
1d. Ribbon		49%	
1e. Cartoon		4%	
1f. Tube		2%	
1g. No preference		0%	
1h. It depends		8%	
2. 3D view preference of a part of a protein	Mult. choice		
2a. Balls		0%	
2b. Balls and sticks		38%	
2c. Sticks		62%	
2d. No preference		0%	
2e. It depends		0%	
3. I often use an option to make a selection of a protein visible	Likert 1-5	3,7	1,0
4. It is easy to recognize residue's structure in a protein	Likert 1-5	2,9	0,9
5. I often use several 3D visualization tools to get more insight into a molecule	Likert 1-5	3,3	1,1
6. I often make only the required information of an interesting residue visible	Likert 1-5	3,8	0,8
7. 3D structure of a protein gives me information about what the function of a residue is	Likert 1-5	3,7	1,0
8. 3D pos. of a residue in a protein in comb. with my backgr. knowledge about the residue always gives me enough info to determine a possible function of a residue	Likert 1-5	3,0	0,9
9. When I know the pos. of a residue, I often use additional resources to gain more information about the protein	Likert 1-5	3,3	1,0
10. When I know the function of an entire protein , I often use additional resources to verify my conclusions	Likert 1-5	3,3	1,0
11. When I know the function of a part of the protein , I often use additional resources to verify my conclusions	Likert 1-5	3,2	1,0
12. I search first for existing info about the protein in sources on the internet before trying to discover more about the protein myself	Likert 1-5	3,3	1,1
13. I often use an option to hide some irrelevant part of the protein	Likert 1-5	3,5	1,0
Part II. SRS & MRS			
1. I often change search options to optimize my search	Likert 1-5	2,7	0,9
2. When I receive results from a tool (e.g. for sequence alignment), I try to change parameters to see how it will influence the results	Likert 1-5	2,7	0,8
3. I often use these cross references to get more information	Likert 1-5	3,7	0,8
4. I use the extended query form	Single choice		
4a. Often		23%	
4b. Sometimes		64%	
4c. Never		13%	
5. When I am not satisfied with the information that I find in the SwissProt database, I use the following additional sources	Mult. choice		
5a. Search engines		18%	
5b. Cross references		28%	
5c. Search manually		15%	
5d. Knowledge from others		19%	
5e. Other		19%	
5f. None		1%	
6. When I search for information about a disease related to a protein, I use the following additional sources to get more information	Mult. choice		
6a. Search engines		30%	

6b. Omim cross references		40%
6c. Other cross references		23%
6d. Other		5%
6e. None		2%
7. When I look for information, the most important to me is	Ranking 1-3	
7a. Detailed answers		19%
7b. Reliable answers		40%
7c. Non-redundant answers		4%
7d. Easy-to-use answers		22%
7e. Well-documented answers		15%
8. When I use cross-references, I use the cross reference according to	Ranking 1-3	
8a. Kind of information		48%
8b. Reliability of the source		31%
8c. Cross references added manually		13%
8d. Cross references added automatically		7%
8e. Other		1%

References

1. Spotfire. *Will better Usability Studies Help Swell Market for Bioinformatics Software?* BioInform 2005, vol. 9 (2)
2. Koua, E., L. and M.J. Kraak. *A usability framework for the design and evaluation of an exploratory geovisualization environment.* in *Eighth International Conference on Information Visualisation (IV'04)*. 2004. Parma, Italy: IEEE.
3. Kosara, R., et al., *Human-centered aspects*, in *Human-Centered Visualization Environments (to be published)*, A. Kerren, A. Ebert, and J. Meyer, Editors. 2006, Springer.
4. Saraiya, P., C. North, and K. Duca. *An Evaluation of Microarray Visualization Tools for Biological Insight.* in *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*. 2004. Washington, DC, USA: IEEE Computer Society.
5. Dunbar, K., *How scientists really reason: Scientific reasoning in real-world laboratories*, in *The Nature of Insight*, R.J. Sternberg and J.E. Davidson, Editors. 1995, The MIT Press: Cambridge, MA. p. 365-395.
6. Bartlett, J.C. and E.G. Toms, *Developing a Protocol for Bioinformatics Analysis: An Integrated Information Behavior and Task Analysis Approach.* Journal of the American Society for Information Science and Technology, 2005. **56**(5): p. 469-482.
7. Dunbar, K., *How Scientists Think: On-Line Creativity and Conceptual Change in Science.* Creative Thought: An Investigation of Conceptual Structures and Processes, 1997.
8. Coughlan, T. and P. Johnson. *Interaction in creative tasks: Ideation, representation and evaluation in composition.* in *CHI 2006*. 2006. Montréal, Canada: ACM.
9. Sundholm, H., H. Artman, and R. Ramberg, *Backdoor Creativity: Collaborative Creativity in Technology Supported Teams*, in *Cooperative systems design: Scenario-based design of collaborative systems*, F. Darses, et al., Editors. 2004, IOS press: Amsterdam. p. 99-114.
10. Ware, C., *Information Visualization: Perception for Design.* Interactive Technologies. 1999, San Francisco, USA: Morgan Kaufmann.
11. Faisal, S., P. Cairns, and B. Craft. *Infovis experience enhancement through mediated interaction.* in *ICMI'05 Workshop on Multimodal Interaction for the Visualization and Exploration of Scientific Data*. 2005. Trento, Italy.
12. Beyer, H., *Contextual Design: Defining Customer-Centered Systems.* 1997, San Francisco, USA: Morgan Kaufmann.
13. Rauwerda, H., et al., *The promise of a Virtual Lab in Drug Discovery.* Drug Discovery Today, 2006. **11**(5-6): p. 228-36.