Pacific Symposium on Biocomputing 13:568-579(2008)

EVIDENCE FOR SHOWING GENE/PROTEIN NAME SUGGESTIONS IN BIOSCIENCE LITERATURE SEARCH INTERFACES

ANNA DIVOLI, MARTI A. HEARST, MICHAEL A. WOOLDRIDGE

School of Information, UC Berkeley
{*divoli,hearst,mikew*}*@.ischool.berkeley.edu*

This paper reports on the results of two questionnaires asking biologists about the incorporation of text-extracted entity information, specifically gene and protein names, into bioscience literature search user interfaces. Among the findings are that study participants want to see gene/protein metadata in combination with organism information; that a significant proportion would like to see gene names grouped by type (synonym, homolog, etc.), and that most participants want to see information that the system is confident about immediately, and see less certain information after taking additional action. These results inform future interface designs.

1. Introduction

Bioinformaticians have developed numerous algorithms for extracting entity and relation information from the bioscience literature, and have developed some very interesting user interfaces for showing this information. However, little research has been done on the *usability* of these systems and how to best incorporate such information into literature search and text mining interfaces. As part of an on-going project to build a highly usable literature search tool for bioscience researchers, we are carefully investigating what kinds of biological information searchers want to see, as well as how they want to see this information presented. We are interested in supporting biologists whose main tasks are biological (as opposed to database curators and bioinformaticians doing text mining) and who presumably do not want to spend a lot of time searching.

We use methods from the field of Human Computer Interaction (HCI) for the careful design of search interfaces. We have already used these methods to develop a novel bioliterature search user interface whose focus is allowing users to search over and view figures and captions[5] (see http://biosearch.berkeley.edu). That interface is based on the observation that many researchers, when assessing a research article, first look at the title, abstract, and figures.

In this paper, we investigate whether or not bioscience literature searchers wish to see related term suggestions, in particular, gene and protein names, in response to their queries.[a] This is one step in a larger investigation in which we plan to assess presentation of other results of text analysis, such as the entities corresponding to diseases, pathways, gene interactions, localization information, function information, and so on.

When it comes to presenting users with the output of text mining programs, the interface designer is faced with an embarrassment of riches. There are many choices of entity and relationship information that can be displayed to the searcher. However, search user interface research suggests that users are quickly overwhelmed when presented with too many options and too much information.

Therefore, our approach is to assess the usability of one feature at a time, see how participants respond, and then test out other features. We focus on gene names here because of their prominent role in the queries devised for the TREC Genomics track[6], and because of their focus in text mining efforts, as seen in the BioCreative text analysis competitions[7]. Thus, this paper assesses one way in which the output of text mining can be useful for bioscience software tools.

In the remainder of this paper, we first describe the user-centered design process and then discuss related work. We then report on the results of two questionnaires. The first asked participants a number of questions about how they search the bioscience literature, including questions about their use of gene names. Among the findings were that participants did indeed want to see suggestions of gene names as part of their search experience. The second questionnaire, building on these results, asked participants to assess several designs for presenting gene names in a search user interface. Finally, we conclude the paper with plans for acting on the results of this study.

2. The User-Centered Design Process

We are following the method of *user-centered design*, which is standard practice in the field of Human-Computer Interaction (HCI)[11]. This method focuses on making decisions about the design of a user interface based on feedback obtained from target users of the system, rather than coding first and evaluating later. First a *needs assessment* is performed in which the designers investigate who the users are, what their goals are, and what tasks they have to complete in order to achieve those goals. The next stage is a *task analysis* in which the designers characterize which steps the users need to take to complete their tasks, decide which user goals they will attempt to support, and then create scenarios which exemplify these tasks being executed by the target user population.

[a] For the remainder of the paper, we will use the term *gene name* to refer to both gene and protein names.

Once the target user goals and tasks have been determined, design is done in a tight evaluation cycle consisting of mocking up prototypes, obtaining reactions from potential users, and revising the designs based on those reactions. This sequence of activities often needs to be repeated several times before a satisfactory design has been achieved. This is often referred to as "discount" usability testing, since useful results can be obtained with only a few participants. After a design is testing well in informal studies, formal experiments comparing different designs and measuring for statistically significant differences can be conducted.

This iterative procedure is necessary because interface design is still more of an art than a science. There are usually several good solutions within the interface design space, and the task of the designers is to navigate through the design space until reaching some local "optimum." The iterative process allows study participants to help the designers make decisions about which paths to explore in that space. Experienced designers often know how to start close to a good solution; less experienced designers need to do more work. Designing for an entirely novel interaction paradigm often requires more iteration and experimentation.

3. Research on Term Suggestions Usability

An important class of query reformulation aids is automatically suggested term refinements and expansions. Spelling correction suggestions are query reformulation aids, but the phrase *term expansion* is usually applied to tools that suggest alternative wordings.

Usability studies are generally positive as to the efficacy of term suggestions when users are not required to make relevance judgements and do not have to choose among too many terms. Those that produce negative results seem to stem from problems with the presentation interface[2]. Interfaces that allow users to reformulate their query by selecting a single term (usually via a hyperlink) seem to fare better. Anick[1] describes the results of a large-scale investigation of the effects of incorporating related term suggestions into a major web search engine. The term suggestion tool, called Prisma, was placed within the Altavista search engine's results page. The number of feedback terms was limited to 12 to conserve space in the display and minimize cognitive load. In a large web-based study, 16% of users applied the Prisma feedback mechanism at least once on any given day. However, effectiveness when measured in the occurrence of search results clicks did not differ between the baseline and the Prisma groups.

In a more recent study, Jansen et al.[9] analyzed 1.5M queries from a log taken in 2005 from the Dogpile.com metasearch engine. The interface for this engine shows suggested additional terms in a box on the righthand side under the heading

"Are you looking for?" Jansen et al. found that 8.4% of all queries were generated by the reformulation assistant provided by Dogpile. Thus, there is evidence that searchers use such term reformulations, although the benefits are as yet unproven.

4. Current Bioliterature Search Interfaces

There are a number of innovative interfaces for analyzing the results of text analysis. The iHOP system[8] converts the contents of PubMed abstracts into a network of information about genes and interactions, displaying sentences extracted from abstracts and annotated with entity information. The ChiliBot[3] system also shows extracted information in the form of relationships between genes, proteins, and keywords. TextPresso[10] uses an ontology to search over the full text of a collection of articles about *C. elegans*, extracting out sentences that contain entities and relations of interest. These systems have not been assessed in terms of usability of their interface or their features.

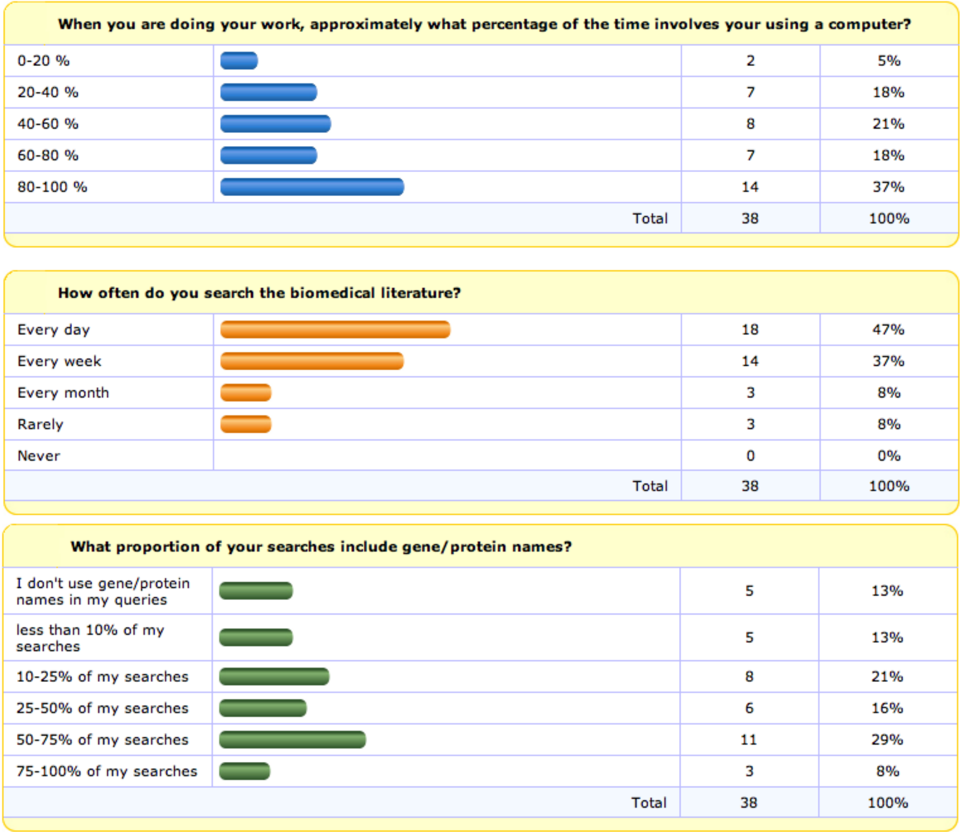
The GoPubMed system[4] shows a wealth of information in search results over PubMed. Most prominent is a hierarchical display of a wide range of categories from the Gene Ontology and MeSH associated with the article. Users may sort search results by navigating in this hierarchy and selecting categories. This interface is compelling, but it is not clear which kinds of information are most useful to show, whether a hierarchy is the best way to show metadata information for grouping search results, and whether or not this is too much information to show. The goal of this paper is to make a start at determining which kinds of information searchers want to see, and how they want to select it.

5. First Questionnaire: Biological Information Preferences

Both studies were administered in the form of an online questionnaire. For the first study, we recruited biosciences researchers from 7 research institutions via email lists and personal contacts. The 38 participants were all from academic institutions (22 graduate students, 6 postdoctoral researchers, 5 faculty, and 5 others), and had a wide range of specialties, including systems biology, bioinformatics, genomics, biochemistry, cellular and evolutionary biology, microbiology, physiology and ecology.

Figure 1 shows the percentage of time each participant uses computers for their work. A surprising 37% say they use computers for 80-100% of the time they are working, although only 6 participants listed bioinformatics as one of their fields. Participants were for the most part heavy users of literature search; 84% said they search biomedical literature either daily or weekly.

We asked participants which existing literature search tools they use, and for



When you are doing your work, approximately what percentage of the time involves your using a computer?			
0-20 %		2	5%
20-40 %		7	18%
40-60 %		8	21%
60-80 %		7	18%
80-100 %		14	37%
	Total	38	100%

How often do you search the biomedical literature?			
Every day		18	47%
Every week		14	37%
Every month		3	8%
Rarely		3	8%
Never		0	0%
	Total	38	100%

What proportion of your searches include gene/protein names?			
I don't use gene/protein names in my queries		5	13%
less than 10% of my searches		5	13%
10-25% of my searches		8	21%
25-50% of my searches		6	16%
50-75% of my searches		11	29%
75-100% of my searches		3	8%
	Total	38	100%

Figure 1. Statistics on computer use, search frequency, and percentage of queries that include gene names.

what percent of their searches. 12 participants (32%) said they use PubMed 80% of the time or more; on average it was used 50% of the time. Google Scholar was used on average 25% of the time; all but 3 participants used it at least some of the time. 6 participants used Ovid at least 5% of the time. The other popular search engine mentioned was the ISI Web of Science, which 9 participants used; 2 said they used it more than 90% of the time. Also mentioned were BIOSIS (3 mentions), Connotea (1), PubMedCentral (1), Google web search (1), and bloglines (1).

Figure 1 shows the responses to a question on what proportion of searches include gene names. 37% of the participants use gene names in 50-100% of their queries. Five participants do not use gene names in their queries; one of these people noted that they use literature search in order to discover relevant genes.

Next, participants answered two detailed questions about what kinds of information they would like to see associated with the gene name from their query. Table 1 shows the averaged scores for responses to the question "When you search for genes/proteins, what type of related gene/protein names would you like a system to suggest?" Participants selected choices from a Likert scale which spanned from 1 ("strongly do not want to see this") to 5 ("extremely important to see this information"), with 3 indicating "do not mind seeing this." (These results are for 33 participants, because the 5 participants who said they do not use gene names in their search were made to automatically skip these questions.) The table below also shows the number of participants who assigned either a 1 or a 2 score, indicating that they do not want to see this kind of information.

Table 1. Averaged scores for responses to the question "When you search for genes/proteins, what type of related gene/protein names would you like a system to suggest?" 1 is "strongly disagree," 5 is "strongly agree."

Related Information Type	Avg. rating	#	(%) selecting 1 or 2
Gene's synonyms	4.4	2	(5%)
Gene's synonyms refined by organism	4.0	2	(5%)
Gene's homologs	3.7	5	(13%)
Genes from the same family: parents	3.4	7	(18%)
Genes from the same family: children	3.6	4	(10%)
Genes from the same family: siblings	3.2	9	(24%)

The next question, "When you search for genes/proteins what other related information would you like a system to return?" used the same rating scale as above. The results are shown in Table 2.

Table 2. Averaged scores for responses to the question "When you search for genes/proteins what other related information would you like a system to return?" using same rating scale as above.

Related Information Type	Avg. rating	#	(%) selecting 1 or 2
Genes this gene interacts with	3.7	4	(10%)
Diseases this gene is associated with	3.4	6	(16%)
Chemicals/drugs this gene is associated with	3.2	8	(21%)
Localization information for this gene	3.7	3	(8%)

When asked for additional information of interest, people suggested: pathways (suggested 4 times), experimental modification, promoter information, lists of organisms for which the gene is sequenced, ability to limit searches to a taxonomic group, protein motifs, hypothesized or known functions, downstream effects and link to a model organism page.

The results of this questionnaire suggest that not only are many biologists heavy users of literature search, but gene names figure prominently in a significant proportion of their searches. Furthermore, there is interest in seeing information associated with gene names. Not surprisingly, the more directly related the information is to the gene, the more participants viewed it favorably. 22 participants said they thought gene synonyms would be extremely useful (i.e., rated this choice with a score of 5). However, as the third columns of the tables show, a notable minority of participants expressed opposition to showing the additional information. In a box asking for general comments, two participants noted that for some kinds of searches, expansion information would be useful, but for others the extra information would be in the way. One participant suggested offering these options at the start of the search as a link to follow optionally. These responses reflect a common view among users of search systems: they do not want to see a cluttered display. This is further warning that one should proceed with caution when adding information to a search user interface.

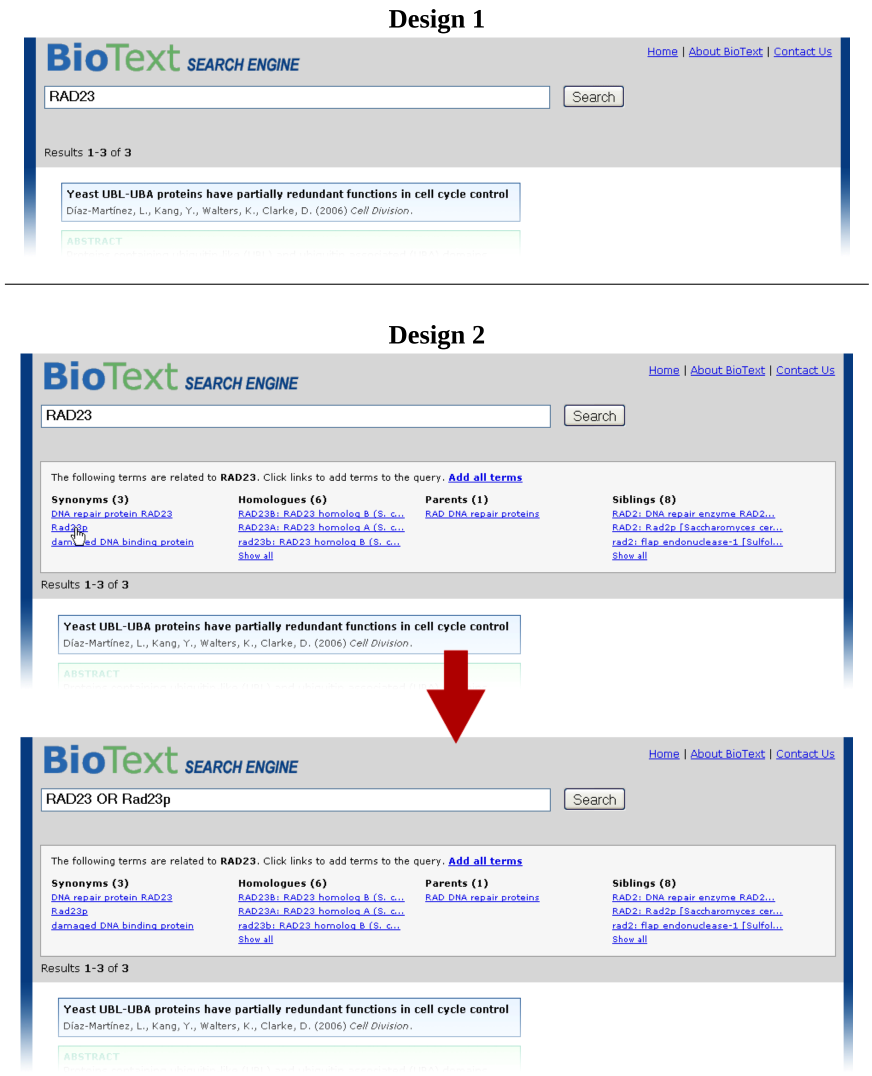
6. Second Questionnaire: Gene/Protein Name Expansion Preferences

6.1. *The Evaluated Designs*

To reproduce what users would see in a Web search interface, four designs were constructed using HTML and CSS, building upon the design used for our group's search engine. To constrain the participants' evaluation of the designs and to focus them on a specific aspect of the interface, static screenshots of just the relevant portion of the search interface were used in the testing. Example interactions with the interface were conveyed using "before" and "after" screenshots of the designs. Limiting the testing to static screenshots decreased the development time required to set up the tests, since we did not need to anticipate the myriad potential interactions between the testers and a live interface. Figures 2–4 show the screenshots seen by the participants for Designs 1–4. Participants were told they were seeing what happened after they clicked on the indicated link, but not what happens to the search results after the new search is executed.

Design 1, which served as the baseline for comparison with the other designs, showed a standard search engine interface with a text box and submit button in the page header. The gene term *"RAD23"* was used as the example search term, with a results summary showing three results returned.

Design 2 added a horizontal box between the search box and the text summary. The box listed possible expansion terms for the original *"RAD23"* query



Design 1

BioText SEARCH ENGINE
Home | About BioText | Contact Us
Search

RAD23

Results 1-3 of 3

Yeast UBL-UBA proteins have partially redundant functions in cell cycle control
Díaz-Martínez, L., Kang, Y., Walters, K., Clarke, D. (2006) *Cell Division*.

Design 2

BioText SEARCH ENGINE
Home | About BioText | Contact Us
Search

RAD23

The following terms are related to **RAD23**. Click links to add terms to the query. **Add all terms**

Synonyms (3)	Homologues (6)	Parents (1)	Siblings (8)
DNA repair protein RAD23	RAD23B: RAD23 homolog B (S. c...	RAD DNA repair proteins	RAD2: DNA repair enzyme RAD2...
Rad23p	RAD23A: RAD23 homolog A (S. c...		RAD2: Rad2p [Saccharomyces cer...
damaged DNA binding protein	rad23b: RAD23 homolog B (S. c...		rad2: flap endonuclease-1 [Sulfol...
	Show all		Show all

Results 1-3 of 3

Yeast UBL-UBA proteins have partially redundant functions in cell cycle control
Díaz-Martínez, L., Kang, Y., Walters, K., Clarke, D. (2006) *Cell Division*.

RAD23 OR Rad23p

The following terms are related to **RAD23**. Click links to add terms to the query. **Add all terms**

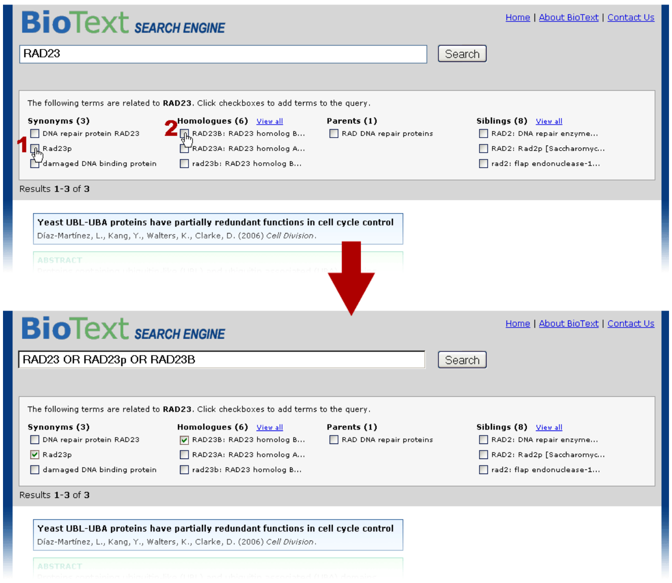
Synonyms (3)	Homologues (6)	Parents (1)	Siblings (8)
DNA repair protein RAD23	RAD23B: RAD23 homolog B (S. c...	RAD DNA repair proteins	RAD2: DNA repair enzyme RAD2...
Rad23p	RAD23A: RAD23 homolog A (S. c...		RAD2: Rad2p [Saccharomyces cer...
damaged DNA binding protein	rad23b: RAD23 homolog B (S. c...		rad2: flap endonuclease-1 [Sulfol...
	Show all		Show all

Results 1-3 of 3

Yeast UBL-UBA proteins have partially redundant functions in cell cycle control
Díaz-Martínez, L., Kang, Y., Walters, K., Clarke, D. (2006) *Cell Division*.

Figure 2. Designs 1 and 2 shown to participants in the second questionnaire.

organized under four categories: synonyms, homologs, parents, and siblings. All the terms were hyperlinked. The "after" screenshot showed the result of clicking a hyperlinked term, which added that term to the query in the text box using an



RAD23

The following terms are related to **RAD23**. Click checkboxes to add terms to the query.

Synonyms (3)
□ DNA repair protein RAD23
□ Rad23p
□ damaged DNA binding protein

Homologues (6)
□ RAD23B: RAD23 homolog B...
□ RAD23A: RAD23 homolog A...
□ rad23b: RAD23 homolog B...

Parents (1)
□ RAD DNA repair proteins

Siblings (8)
□ RAD2: DNA repair enzyme...
□ RAD2: Rad2p [Saccharomyc...
□ rad2: flap endonuclease-1...

Results **1-3** of **3**

Yeast UBL-UBA proteins have partially redundant functions in cell cycle control
Díaz-Martínez, L., Kang, Y., Walters, K., Clarke, D. (2006) *Cell Division*.

RAD23 OR RAD23p OR RAD23B

The following terms are related to **RAD23**. Click checkboxes to add terms to the query.

Synonyms (3)
□ DNA repair protein RAD23
☑ Rad23p
□ damaged DNA binding protein

Homologues (6)
☑ RAD23B: RAD23 homolog B...
□ RAD23A: RAD23 homolog A...
□ rad23b: RAD23 homolog B...

Parents (1)
□ RAD DNA repair proteins

Siblings (8)
□ RAD2: DNA repair enzyme...
□ RAD2: Rad2p [Saccharomyc...
□ rad2: flap endonuclease-1...

Results **1-3** of **3**

Yeast UBL-UBA proteins have partially redundant functions in cell cycle control
Díaz-Martínez, L., Kang, Y., Walters, K., Clarke, D. (2006) *Cell Division*.

Figure 3. Design 3 shown to participants in the second questionnaire.

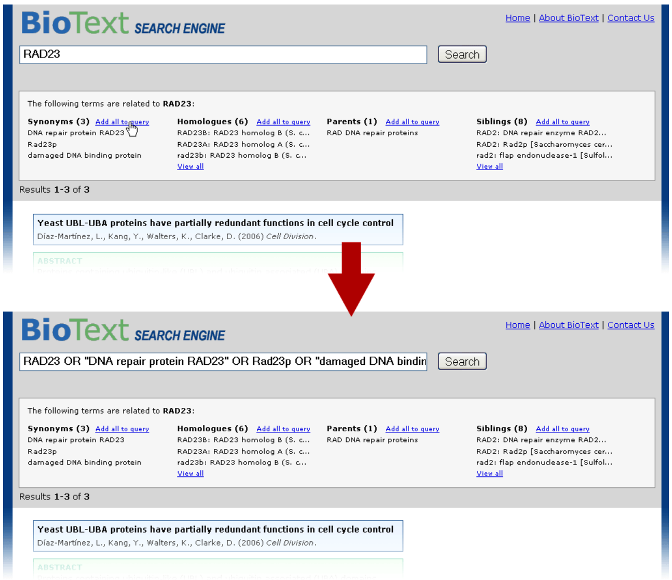
OR operator.

Design 3 had a similar layout except that instead of having hyperlinked expansion terms, each expansion term was paired with a checkbox. The terms were organized beneath the same four categories. The "after" screenshot showed that by clicking a checkbox, a user could add the term to the original query.

Design 4 showed a box of plain text expansion terms that were neither hyperlinked nor paired with checkboxes. In this design, each category term had an "Add all to query" link next to it for adding all of a category's terms at once. The "after" screenshot showed the result of clicking a hyperlink, with multiple terms ORed to the original query.

6.2. *Results*

Nineteen people completed the questionnaire. Nine of those who filled out the first questionnaire and who indicated that (a) they were interested in seeing gene/protein names in search results and (b) they were willing to be contacted for a second questionnaire participated in this followup study. Ten additional participants were recruited by emailing colleagues and asking them to forward the



RAD23

The following terms are related to **RAD23**:

Synonyms (3)
DNA repair protein RAD23
Rad23p
damaged DNA binding protein

Homologues (6)
RAD23B: RAD23 homolog B (S. c...
RAD23A: RAD23 homolog A (S. c...
rad23b: RAD23 homolog B (S. c...

Parents (1)
RAD DNA repair proteins

Siblings (8)
RAD2: DNA repair enzyme RAD2...
RAD2: Rad2p [Saccharomyces cer...
rad2: flap endonuclease-1 [Sulfol...

Results **1-3** of **3**

Yeast UBL-UBA proteins have partially redundant functions in cell cycle control
Díaz-Martínez, L., Kang, Y., Walters, K., Clarke, D. (2006) *Cell Division*.

RAD23 OR "DNA repair protein RAD23" OR Rad23p OR "damaged DNA bindin

The following terms are related to **RAD23**:

Synonyms (3)
DNA repair protein RAD23
Rad23p
damaged DNA binding protein

Homologues (6)
RAD23B: RAD23 homolog B (S. c...
RAD23A: RAD23 homolog A (S. c...
rad23b: RAD23 homolog B (S. c...

Parents (1)
RAD DNA repair proteins

Siblings (8)
RAD2: DNA repair enzyme RAD2...
RAD2: Rad2p [Saccharomyces cer...
rad2: flap endonuclease-1 [Sulfol...

Results **1-3** of **3**

Yeast UBL-UBA proteins have partially redundant functions in cell cycle control
Díaz-Martínez, L., Kang, Y., Walters, K., Clarke, D. (2006) *Cell Division*.

Figure 4. Design 4 shown to participants in the second questionnaire.

request to biologists. Thus, the results are biased towards people who are interested in search interfaces and their improvement.

Again, participants were from several academic institutions (4 graduate students, 7 postdoctoral researchers, 3 faculty, and 5 other researchers). Their areas of interest/specialization included molecular toxicology, evolutionary genomics, chromosome biology, plant reproductive biology, cell signaling networks, and computational biology more generally. The distribution of usage of genes in searches was similar to that of the first questionnaire.

One question asked the participants to rank-order the designs. There was a clear preference for the expansion terms over the baseline, which was the lowest ranked for 15 out of 19 participants. Table 3 shows the results, with Design 3 most favored, followed by Designs 4 and 2, which were similarly ranked.

In the next phase of questions, one participant indicated they would not like to see gene names, and so automatically skipped the questions. Of the remaining 18 participants, when asked to indicate a preference for clicking on hyperlinks versus checkboxes for adding gene names to the query, 10 participants (56%) selected checkboxes and 6 (33%) selected hyperlinks (one suggested a “select all”

Table 3. Design Preferences.

	# participants who rated Design 1st or 2nd	% participants who rated Design 1st or 2nd	Avg. rating (1=low, 4=high)
Design 3	15	79%	3.3
Design 4	10	53%	2.6
Design 2	9	47%	2.5
Design 1	0	0%	1.6

option above each group for the checkboxes). When asked to indicate whether or not they would like to see the organisms associated with each gene name, 16 out of 18 participants said they would like the organism information to be directly visible, showing the organism either alongside (11) or grouping the gene names by organism (5). Two were undecided.

When asked how gene names should be organized in the display, 9 preferred them to be grouped under type (synonyms, homologs, etc). The other participants were split between preferences for showing the information grouped by organism name, grouped by more generic taxonomic information, or not grouped but shown alphabetically or by frequency of occurrence in the collection.

Participants were also asked if they prefer to select each gene individually (2), whole groups of gene names with one click (3), or to have the option to chose either individual names or whole groups with one click (13).

Finally, they were asked if they prefer the system to suggest only names that it is highly confident are related (8), include names that it is less confident about (0), or include names that it is less confident about under a “show more” link (8). In the open comments field, one participant stated that the system should allow the user to choose among these, and another wrote something we could not interpret. These attitudes echo the finding that high-scoring systems in the TREC genomics track[6] often used principled gene name expansion.

7. Conclusions and Future Work

This study addresses the results of the first steps of user-centered design for development of a literature search interface for biologists. Our needs assessment has revealed a strong desire for the search system to suggest information closely related to gene names, and some interest in less closely related information as well. Our task analysis has revealed that most participants want to see organism names in conjunction with gene names, a majority of participants prefer to see term suggestions grouped by type, and participants are split in preference between single-click hyperlink interaction and checkbox-style interaction. The last point suggests that we experiment with hybrid designs in which only hyperlinks are used, but an additional new hyperlink allows for selecting all items in a group. Another hybrid to evaluate would have checkboxes for the individual terms and a link that immediately adds all terms in the group and executes the query.

The second questionnaire did not ask participants to choose between seeing information related to genes and other kinds of metadata such as disease names. Adding additional information will require a delicate balancing act between usefulness and clutter. Another design idea would allow users to collapse and expand term suggestions of different types; we intend to test that as well.

Armed with these results, we have reason to be confident that the designs will be found usable. Our next steps will be to implement prototypes of these designs, ask participants to perform queries, and contrast the different interaction styles.

Acknowledgements: We thank the survey participants for their contributions to this work. This research was supported in part by NSF DBI-0317510.

References

1. P. Anick. Using terminological feedback for web search refinement: a log-based study. *Proceedings of SIGIR 2003*, pages 88–95, 2003.
2. P. Bruza, R. McArthur, and S. Dennis. Interactive Internet search: keyword, directory and query reformulation mechanisms compared. *Proceedings of SIGIR 2000*, pages 280–287, 2000.
3. H. Chen and B.M. Sharp. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, 5(147), 2004.
4. A. Doms and M. Schroeder. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, 33(1):W783–W786, 2005.
5. M.A. Hearst, A. Divoli, J. Ye, and M.A. Wooldridge. Exploring the efficacy of caption search for bioscience journal search interfaces. *Biological, translational, and clinical language processing*, pages 73–80, 2007.
6. W. Hersh, A. Cohen, J. Yang, R.T. Bhupatiraju, P. Roberts, and M. Hearst. TREC 2005 Genomics Track Overview. *the Fourteenth Text Retrieval Conference*, 2005.
7. L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(S1), 2005.
8. R. Hoffmann and A. Valencia. A gene network for navigating the literature. *Nature Genetics*, 36(7):664, 2004.
9. B.J. Jansen, A. Spink, and S. Koshman. Web searcher interaction with the Dogpile.com metasearch engine. *Journal of the American Society for Information Science and Technology*, 58(5):744–755, 2007.
10. H.M. Müller, E.E. Kenny, and P.W. Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309, 2004.
11. B. Shneiderman and C. Plaisant. *Designing the user interface: strategies for effective human-computer interaction, 4/E*. Addison-Wesley, Reading, MA, 2004.