
Intelligent Information Management
Targeted Competition Framework
ICT-2011.4.4(d)

Project **FP7-318652 / BioASQ**

Deliverable **D3.4**

Distribution **Public**


http://www.bioasq.org

# Tutorials and Guidelines

Prodromos Malakasiotis, Ion Androutsopoulos, Yannis Almirantis, Dimitris Polychronopoulos and Ioannis Pavlopoulos

Status: Final (Version 1.0)

January 2013

## Project

| | |
|---|---|
| Project ref.no. | FP7-318652 |
| Project acronym | BioASQ |
| Project full title | A challenge on large-scale biomedical semantic indexing and question answering |
| Porject site | http://www.bioasq.org |
| Project start | October 2012 |
| Project duration | 2 years |
| EC Project Officer | Martina Eydner |

## Deliverable

| | |
|---|---|
| Deliverabe type | Report |
| Distribution level | Public |
| Deliverable Number | D3.4 |
| Deliverable title | Tutorials and Guidelines |
| Contractual date of delivery | M4 (January 2013) |
| Actual date of delivery | January 2013 |
| Relevant Task(s) | WP3/Task 3.4 |
| Partner Responsible | AUEB-RC |
| Other contributors | NCSR “D” |
| Number of pages | 25 |
| Author(s) | Prodromos Malakasiotis, Ion Androutsopoulos, Yannis Almirantis, Dimitris Polychronopoulos and Ioannis Pavlopoulos |
| Internal Reviewers | George Balikas, Patric Gallinari |
| Status & version | Final |
| Keywords | benchmark datasets, biomedical experts, guidelines, tutorial, annotation tool |

# Executive Summary

This deliverable comprises the tutorial and guidelines that will be provided to the team of biomedical experts to help them create the questions, reference answers, and other supportive information that will be used in the benchmark dataset of the first BioASQ challenge. The guidelines provide directions regarding the number and types of questions to be created by the experts, the information sources the experts should consider and how to use them, the types and sizes of the reference answers and the other supportive information the experts should provide etc. The annotation tool of deliverable D3.3 was designed to help the biomedical experts follow the same guidelines within a unified and easy to use Web interface that provides access to all the necessary resources, allows the questions, reference answers, and supportive information to be edited, saved etc. A tutorial illustrating the usage of the annotation tool is included in this deliverable, and will be provided to the biomedical expert team, along with access to the tool and the guidelines. More technical information about the annotation tool can be found in deliverable D3.3.

# Contents

# List of Figures

# 1

# Introduction

This deliverable comprises the tutorial and guidelines that will be provided to the BioASQ team of biomedical experts to help them create the questions, reference answers, and other supportive information that will be used in the benchmark dataset of the first BioASQ challenge.

The guidelines provide directions regarding the number and types of questions to be created by the experts, the information sources the experts should consider and how to use them, the types and sizes of the reference answers and the other supportive information the experts should provide etc. The annotation tool of deliverable D3.3 was designed to help the biomedical experts follow the same guidelines within a unified and easy to use Web interface that provides access to all the necessary resources, allows the questions, reference answers, and supportive information to be edited, saved etc. A tutorial illustrating the usage of the annotation tool is included in this deliverable, and will be provided to the biomedical expert team, along with access to the tool and the guidelines. More technical information about the annotation tool can be found in deliverable D3.3.

Chapters 2 and 3 below present the guidelines and the tutorial, respectively, that will be provided to the biomedical expert team. The tutorial, which will also be available as a slide-show presentation, presupposes that the experts have studied the guidelines. The guidelines were developed by consulting the biomedical expert team and conducting a pilot study, which is discussed in Appendix A. We would also like to thank the several members of the BioASQ advisory board who provided information on the datasets and tools of other previous related competitions.

# 2

# Benchmark Creation Guidelines

Each biomedical expert should formulate *at least* 30 English questions, reflecting real-life information needs encountered during his/her work (e.g., in research or diagnosis). Each question should be stand-alone, i.e., it should not presuppose that any other questions have been asked; for example, it should not contain any pronouns referring to entities mentioned in other questions. For each question, the expert is also expected to provide an answer and other supportive information, as explained below.

To formulate each question and to provide the corresponding answer and supportive information, the expert should follow the following steps. An *annotation tool* will be made available to help the experts follow these steps, and a tutorial showing how to use the tool is provided in Chapter 3.

**Step 1: Question formulation.** Formulate an English stand-alone question reflecting real-life information needs. *At least 5 questions of each one of the following four categories* should be formulated by each biomedical expert; more than 5 questions will have to be formulated for some of the four categories, since *a total of at least 30 questions* is required.

**Yes/no questions:** These are questions that, strictly speaking, require either a "yes" or a "no" as an answer, though of course in practice a longer answer providing additional information that supports the "yes" or "no" will often be desirable. For example, "*Do CpG islands colocalise with transcription start sites?*" is a yes/no question.

**Factoid questions:** These are questions that, strictly speaking, require a particular entity (e.g., a disease, drug, or gene) as an answer, though again a longer answer providing additional supportive information may be desirable in practice. For example, "*Which virus is best known as the cause of infectious mononucleosis?*" is a factoid question.

**List questions:** These are questions that, strictly speaking, require a *list* of entities (e.g., a list of genes) as an answer; again, in practice additional supportive information may be desirable. For example, "*Which are the Raf kinase inhibitors?*" is a list question.

**Summary questions:** These are questions that do not belong in any of the previous categories and can only be answered by producing a short text summarizing the most prominent relevant information. For example, "*What is the treatment of infectious mononucleosis?*" is a

summary question. When formulating summary questions, the experts should aim at questions that they can answer (possibly after consulting the literature) in a satisfactory manner by writing a one-paragraph summary intended to be read by other experts of the same field.

In all four categories of questions, the experts should aim at questions that when converted to PUBMEDCENTRAL queries, as discussed below, retrieve approximately 10–60 articles (or abstracts). Questions for which there are controversial or no answers in the literature should be avoided.

**Step 2: Relevant terms.** Form a set of terms that are relevant to the question of Step 1. The set of relevant terms may include terms that are already mentioned in the question, but it may also include synonyms of the question terms, closely related broader and narrower terms etc. For the question "*Do CpG islands colocalise with transcription start sites?*", the set of relevant terms would most probably include the question terms "*CpG Island*" and "*transcription start site*", and possibly also other terms.

**Step 3: Information retrieval.** Facilities will be provided to formulate a query (Boolean or simple bag of terms) involving the relevant terms of Step 2, as well as to retrieve articles from PUBMEDCENTRAL that satisfy the query (or abstracts, when only abstracts are available). The query can be enriched with the advanced search tags of PUBMEDCENTRAL.[1] Facilities will also be provided to execute the query against biomedical terminology banks, databases, and knowledge bases, in order to obtain possibly relevant *concepts* (e.g., MESH headings) and relations (e.g., a database may show that a particular disease is known to cause a particular symptom). Relations retrieved from databases and knowledge bases will be shown in the annotation tool as pseudo-natural language statements, hereby called simply *statements*; hence, the experts do not need to be familiar with how information is actually represented in the databases and knowledge bases. Note that when retrieving concepts and statements, advanced search tags are ignored. Furthermore, when retrieving statements, Boolean operators are also ignored, i.e., Boolean queries are turned into bag of terms queries.

Returning to the example question "*Do CpG islands colocalise with transcription start sites?*" of Step 1, a possible Boolean query involving the relevant terms of Step 2 might be "*CpG Island*" *AND* "*transcription start site*". The concepts, articles, and statements retrieved by this query are shown below; we only show the titles of the articles to save space, but the annotation tool will allow the experts to view the entire articles or their abstracts (when only abstracts are available).[2] Shown in brackets are the names of the resources the concepts come from.

Retrieved concepts:

1. *"Transcription Initiation Site"* (MESH)
2. *"Factor VIII intron 22 protein"* (UNIPROT)
3. *"Factor VIII intron 22 protein"* (UNIPROT)
4. *"CpG Islands"* (MESH)
5. *"regulation of transcription, start site selection"* (GENE Ontology)
6. *"hypermethylation of CpG island"* (GENE Ontology)
7. *"hypomethylation of CpG island"* (GENE Ontology)

[1] See `http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Search_Field_Descrip` for a detailed description of the tags.

[2] More concepts are actually retrieved; we only show the first 10 to save space.

8. *"DNA-dependent transcriptional start site selection"* (GENE Ontology)
9. *"Cyclic 2,3-diphosphoglycerate synthetase"* (UNIPROT)
10. *"Cyclic 2,3-diphosphoglycerate synthetase"* (UNIPROT)

Retrieved articles (only titles shown here):

1. *"Putative Zinc Finger Protein Binding Sites Are Over-Represented in the Boundaries of Methylation-Resistant CpG Islands in the Human Genome"*
2. *"CpG Islands: Starting Blocks for Replication and Transcription"*
3. *"Periodicity of SNP distribution around transcription start sites"*
4. *"Comprehensive analysis of the base composition around the transcription start site in Metazoa"*
5. *"DBTSS: DataBase of Human Transcription Start Sites, progress report 2006"*
6. *"Assessment of clusters of transcription factor binding sites in relationship to human promoter, CpG islands and gene expression"*
7. *"CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences"*
8. *"CpG islands in vertebrate genomes"*
9. *"Dynamic usage of transcription start sites within core promoters"*
10. *"Boosting with stumps for predicting transcription start sites"*

Retrieved statements:

1. *"Methyl-cpg-binding domain protein 2's specific function is binds cpg islands in promoters where the dna is methylated at position 5 of cytosine within cpg dinucleotides. binds hemi-methylated dna as well. recruits histone deacetylases and dna methyltransferases. acts as transcriptional repressor and plays a role in gene silencing. isoform 1 may enhance the activation of some unmethylated camp-responsive promoters. reports about dna demethylase activity of isoform 2 are contradictory."*

**Step 4: Selection of concepts, articles, statements.** *All* the concepts of Step 3 that best characterise the question of Step 1 should be selected at this step. Also, *all* the articles of Step 3 that are *possibly relevant* to the question should be selected. By 'possibly relevant' we mean articles that the expert would want to read more carefully in practice, to check if they contain information that is useful to answer the question. At this step, the expert is only expected to skim through the retrieved articles (or their abstracts) to figure out if they are possibly relevant. Finally, *every* statement of Step 3 that provides information that is useful to answer the question should be selected, even if the statement does not provide on its own all of the information that is needed to answer the question. In our example, the following concepts, documents, and statements might be selected:

Selected concepts:

1. *"Transcription Initiation Site"* (MESH)
4. *"CpG Islands"* (MESH)
5. *"regulation of transcription, start site selection"* (GENE Ontology)
6. *"hypermethylation of CpG island"* (GENE Ontology)
7. *"hypomethylation of CpG island"* (GENE Ontology)
8. *"DNA-dependent transcriptional start site selection"* (GENE Ontology)

Selected articles (only titles shown here):

2. *"CpG Islands: Starting Blocks for Replication and Transcription"*
4. *"Comprehensive analysis of the base composition around the transcription start site in Metazoa"*
5. *"DBTSS: DataBase of Human Transcription Start Sites, progress report 2006"*
7. *"CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences"*
8. *"CpG islands in vertebrate genomes"*
9. *"Dynamic usage of transcription start sites within core promoters"*
10. *"Boosting with stumps for predicting transcription start sites"*

Selected statements:

1. *"Methyl-cpg-binding domain protein 2's specific function is binds cpg islands in promoters where the dna is methylated at position 5 of cytosine within cpg dinucleotides. binds hemi-methylated dna as well. recruits histone deacetylases and dna methyltransferases. acts as transcriptional repressor and plays a role in gene silencing. isoform 1 may enhance the activation of some unmethylated camp-responsive promoters. reports about dna demethylase activity of isoform 2 are contradictory."*

**Step 5: Text snippet extraction.** At this stage, the expert should read (or skim through more carefully) the set of possibly relevant articles selected during Step 4. *Every* text snippet (piece of text) that provides information that is useful to answer the question of Step 1 should be extracted, even if the snippet on its own does not provide all of the information that is needed to answer the question. The experts should avoid including in the extracted snippets long pieces of text that do not provide useful information; for example, if only a sentence (or part of a sentence) of a paragraph provides useful information, only that sentence (or part of that sentence) should be extracted as a snippet. On the other hand, the experts should not spend too much time trying to decide exactly where each extracted snippet should start or end; only approximate snippet boundaries are needed. If there are multiple snippets that provide the same (or almost the same) useful information (in the same article or in different articles), *all* of them should be extracted, not just one of them. Snippets can be easily extracted using the annotation tool, much as one might highlight snippets that provide useful information when reading an article. In our example, the following snippets might be extracted. The numbers in square brackets point to the articles of Step 4 the snippets were extracted from.

- *"A common explanation for the G+C rise that is seen here in the mammalian profile in the proximity of the TSS is the presence of CpG islands,"* [4]
- *"Above we have made the remark that the G+C rise in mammals and maybe generally in vertebrates is probably caused by the higher number of CpG dinucleotides in the promoter region."* [4]
- *"This could mean that there is some DNA methylation and some CpG over-representation around TSS but not as much as in human."* [4]
- *"The results for Fugu (Fig. 4C,4D) show that some genes could have CpG islands (Fig. 4D) since for those the nucleotide composition is similar to the mammalian profiles."* [4]
- *"Nucleotide composition and gene expressionIt is generally known that the presence of a CpG island around the TSS is related to the expression pattern of the gene. Unmethylated DNA can have an open chromatin structure that facilitates the interaction of transcription*

*factors with the promoter region [15]. Housekeeping genes (HK genes), which are transcribed in all somatic cells and under all circumstances (and thus should be easily activated) frequently have a CpG island in their promoter region [16,17].”* [4]

- *“CpG islands are good markers of some classes of genes because they are often linked to the promoters of those genes”* [5]
- *“In most cases, CpG islands escape DNA methylation, which suppresses gene expression in general, in almost every tissue [10] and function as part of the gene promoter [11].”* [5]
- *“In the human genome, CpG-rich promoters or CpG island promoters are dominant, occurring more than twice as often as CpG-poor promoters”* [5]
- *“Currently, the presence of CpG islands in invertebrate animals is unclear.”* [5]
- *“It is well known that the enrichment of the CpG dinucleotides in CpG island promoters is maximum in TSSs [12,13], so TSSs constitute candidate regions in which CpG island promoters or CpG island-like sequences might occur in the invertebrate genome.”* [5]
- *“The CpG-rich promoters can be considered to contain a CpG island.”* [5]
- *“his observation led to the hypothesis that human CpG-poor promoters emerged with the deamination of methylated CpG dinucleotides in CpG island promoters”* [5]
- *“Our results confirmed that the ascidian promoters tended to have high CpG score and G+C contents around TSS, as was observed in the human promoters.”* [5]
- *“Although the ascidian TSSs exhibited quite high CpG score, this fact does not necessarily mean that they have high frequency of the CpG dinucleotide”* [5]
- *“ascidian promoters tended to exhibit high CpG scores”* [5]
- *“CpG island promoters must have appeared in an early stage of vertebrate evolution”* [5]
- *“The sequences near TSSs tend to exhibit high CpG score and high G+C content, but their level and extent are actually restricted.”* [5]
- *“Considering that 67.5% of responsive genes have CpG islands,”* [7]
- *“We found that more than a third (33.4-34.1%) of these tissue-specific genes had CpG islands”* [7]
- *“another observation that 24% of brain-specific promoters have CpG islands”* [7]
- *“CGIs often extend into downstream transcript regions. This provides an explanation for the observation that the exon at the 5’ end of the transcript, flanked with the transcription start site, shows a remarkably higher CpG density than the downstream exons”* [8]
- *“Genes with a CGI in their promoter tended to be regulated by H3K36me3 rather than nucleosomes or CpG methylation, probably for efficient transcription elongation”* [8]
- *“CGIs and NFRs tend to coexist in some promoters, together marking an active chromatin configuration”* [8]
- *“CpG methylation is proposed to cooperate with nucleosomes and H3K36me3 to differentially regulate the elongation of pol II.”* [8]
- *“These associations are consistent with the previous finding that broad tag clusters are associated with CpG islands”* [9]
- *“An interpretation of this fine-grained tissue specificity is that the differential methylation of each CpG dinucleotide affects the transcription machinery, and results in different specificities without a clear positional bias”* [89]

– *"Although there has been much success in locating the TSSs for CpG-related promoters, the performance for non-CpG-related promoters (about 25% of known genes) is still not satisfactory because of the diverse nature of vertebrate promoter sequences"* [10]

**Step 6: Query revision.** If the expert believes that the snippets and statements gathered during Steps 4 and 5 do not provide enough information to answer the question, the terms of Step 2 and the query of Step 3 should be revised, for example using more or different terms. The process will then continue from Step 3, i.e., the revised query will be used to perform a new search, which may produce different concepts, articles, and statements; the expert will again select (in Step 4) concepts, articles, and statements among those retrieved, and then snippets (in Step 5). The annotation tool provides facilities that allow the concepts, articles, and statements that the expert has already selected (before performing a new search) to be saved, along with the snippets the expert has already extracted. The query can be revised several times, until the expert feels that the gathered information is sufficient to answer the question. If despite revising the query the expert feels that the gathered information is insufficient, or if there seem to be controversial answers, the question should be discarded.

**Step 7: Exact answer.** At this step, the expert should provide what we call an *exact answer* for the question of Step 1. For a yes/no question, the exact answer should be simply "yes" or "no". For a factoid question, the exact answer should be the name of the entity (e.g., gene, disease) seeked by the question; if the entity has several names, the expert should provide, to the extent possible, all of its names, as explained in the tutorial of Chapter 3. For a list question, the exact answer should be a list containing the entities seeked by the question; if a member of the list has several names, the expert should provide, to the extent possible, all of the member's names, again as explained in the tutorial of Chapter 3. For a summary question, the exact answer should be left blank. The exact answers of yes/no, factoid, and list questions should be based on the information of the statements and text snippets that the expert has selected and extracted in Steps 4 and 5, respectively, rather than, for example, personal experience.

**Step 8: Ideal answer.** At this step, the expert should formulate what we call an *ideal answer* for the question of Step 1. The ideal answer should be a one-paragraph text that answers the question of Step 1 in a manner that the expert finds satisfactory. The ideal answer should be written in English, and it should be intended to be read by other experts of the same field. For the example yes/no question "*Do CpG islands colocalise with transcription start sites?*", an ideal answer might be the following:

*"Yes. It is generally known that the presence of a CpG island around the TSS is related to the expression pattern of the gene. CGIs (CpG islands) often extend into downstream transcript regions. This provides an explanation for the observation that the exon at the 5' end of the transcript, flanked with the transcription start site, shows a remarkably higher CpG density than the downstream exons."*

The ideal answer should be based on the information of the statements and text snippets that the expert has selected and extracted in Steps 4 and 5, respectively, rather than, for example, personal experience. The experts, however, are allowed (and should) rephrase or shorten the statements and snippets, order or combine them etc., in order to make the ideal answer more concise and easier to read etc.

Notice that in the example above, the ideal answer is longer than the exact one ("yes"), and that the idea answer provides additional information supporting the exact answer. If the expert feels that

the exact answer of a yes/no, factoid, or list question is sufficient and no additional information needs to be reported, the ideal answer can be the same as the exact answer. For summary questions, an ideal answer must always be provided.

# 3

# Annotation Tool Tutorial

The biomedical experts will be assisted in creating the benchmark sets (questions, answers, and supportive information) by an annotation tool. The annotation tool can be used via a Web interface, which is available at: `http://at.bioasq.org/`. This chapter demonstrates the usage of the annotation tool, assuming that the reader has already studied the guidelines of Chapter 2. More technical information about the annotation tool can be found in deliverable D3.3.

## 3.1 Registration and log-in

Each biomedical expert should first register to use the annotation tool. To register, click on the "Register" button of the initial page (Figure 3.1) of the annotation tool (`http://at.bioasq.org/`). A form (Figure 3.2) will appear, where each biomedical expert should fill in his/her e-mail address, name (first name followed by last name), and a desired password, to be re-typed in the "password repeat" field. Clicking on the "Register!" button of the registration form (Figure 3.2) submits the registration request. A confirmation e-mail message will be sent to the expert. The e-mail message will include a link that the expert should click on to complete the registration process. Once registered, the expert can log in to the annotation tool by filling in his/her e-mail address and password (the ones entered during registration) and clicking on the the "Login" button of the annotation tool initial page (Figure 3.3). Experts who have forgotten their passwords should click on the "Forgot your password?" button (Figure 3.3) to receive further instructions.

## 3.2 Question formulation

Having logged in, the expert can create a new question by clicking on the "+New" button (Figure 3.4). A form will then appear (Figure 3.5), where the expert can fill in the question (in English) and select its type ("yes/no" question, factoid question, list question, or summary question). Consult Step 1 of the guidelines of Chapter 2 for more information on the types of questions.

After filling in the question and selecting its type, the expert should click on the "OK" button (Figure 3.5) to return to the previous page. There (Figure 3.4) the expert can select from the drop-down menu

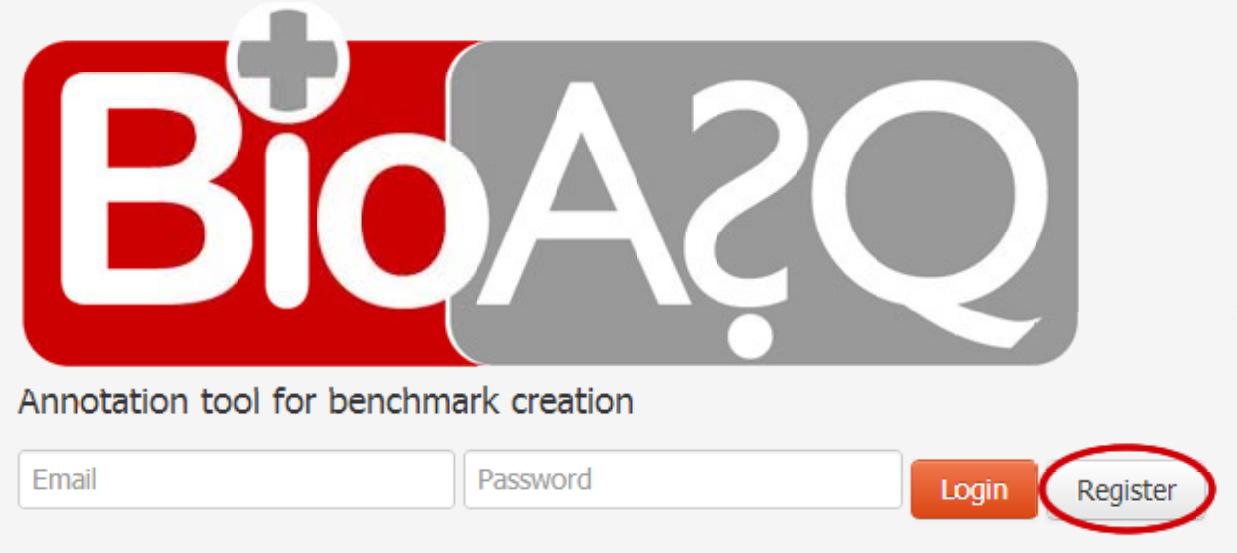


Figure 3.1: Initial page of the annotation tool.

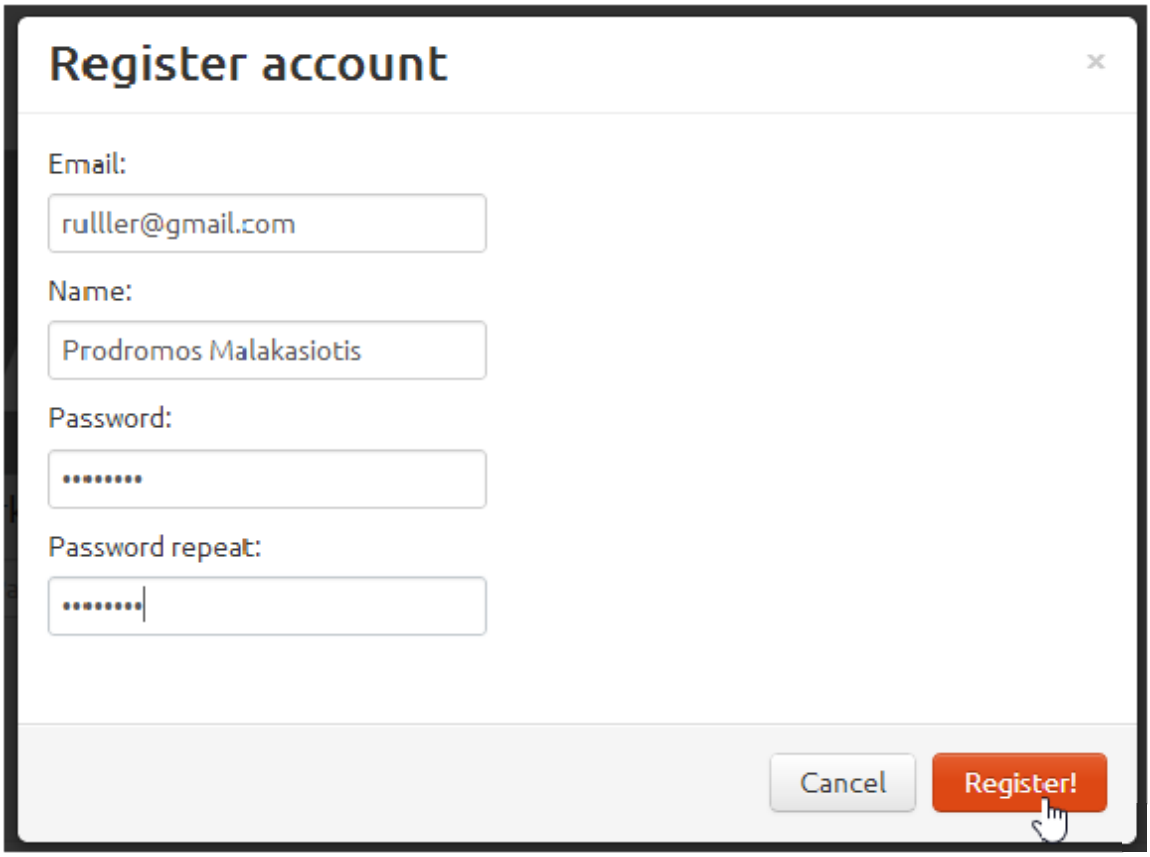


Figure 3.2: Registration form of the annotation tool.

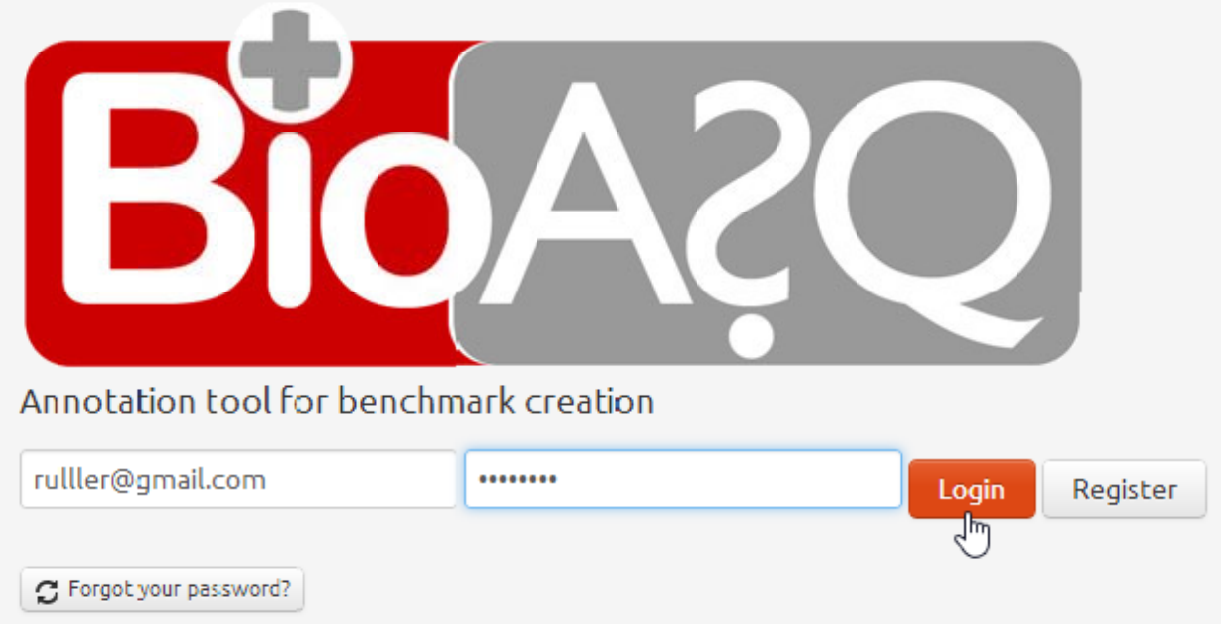


Figure 3.3: Logging in to the annotation tool.

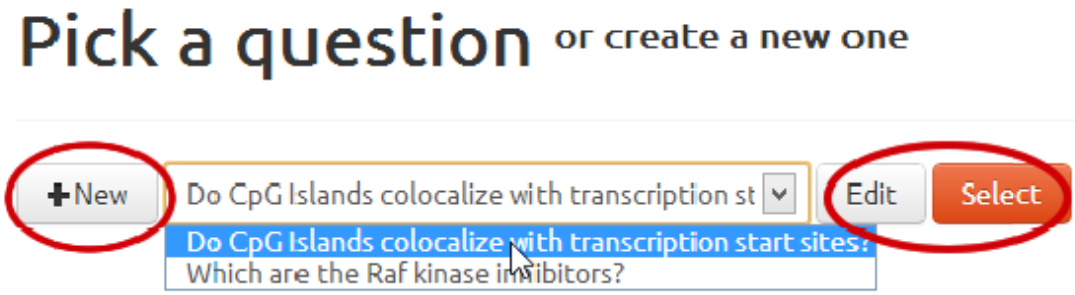


Figure 3.4: Creating a new question or selecting a previously created one.

either the newly created question or a previous question he/she has created, in order to perform further work on that question. Having selected a question, the expert should click on the "Select" button (Figure 3.4) to proceed.

The expert is also provided with the option to edit or delete a question. This can be achieved by clicking the "Edit" button (Figure 3.5). A form then appears (Figure 3.6), where the expert can edit a question, change its type or delete it.

## 3.3 Relevant terms and information retrieval

Having selected a question to work with, the expert can proceed to formulate a query involving terms that are relevant to the question, as discussed in Steps 2 and 3 of the guidelines of Chapter 2. The query has to be entered in the "Query..." text box of Figure 3.7. It can be a "bag-of-words" query or a Boolean query. A bag-of-words query is simply a set of terms, as in the following example:

```
"di-glycine signature" Trypsin human
```

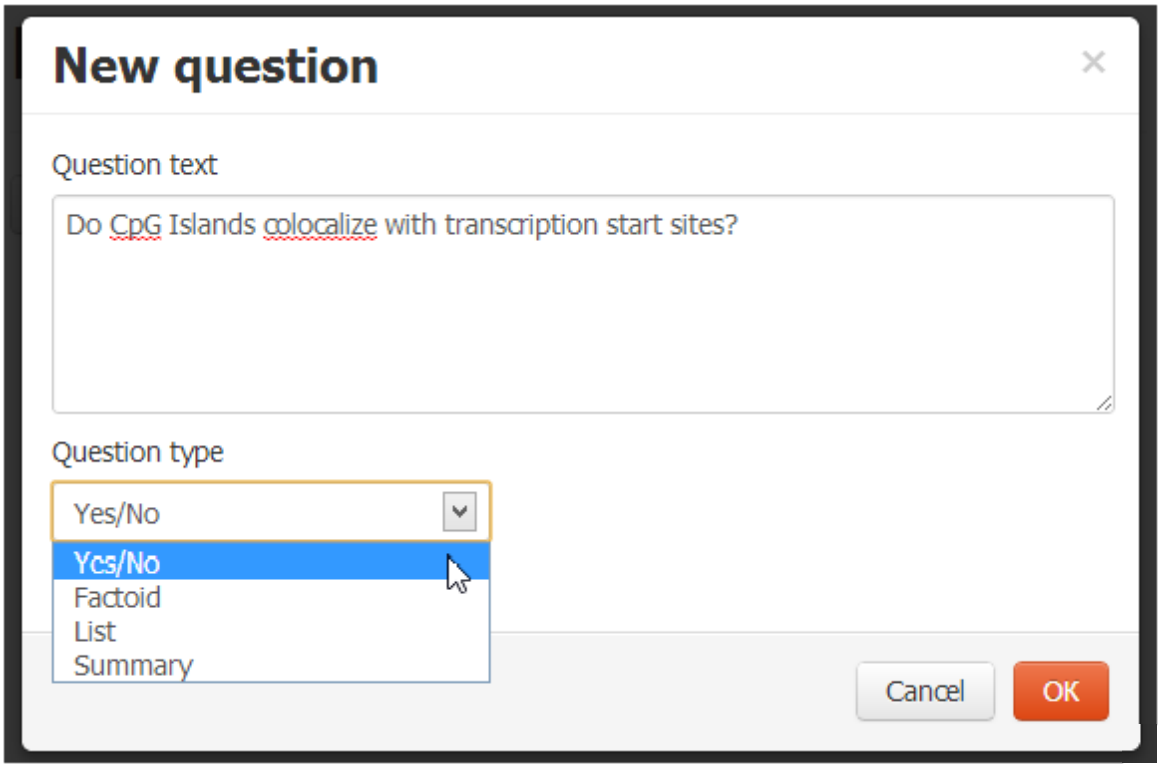


Figure 3.5: New question form.

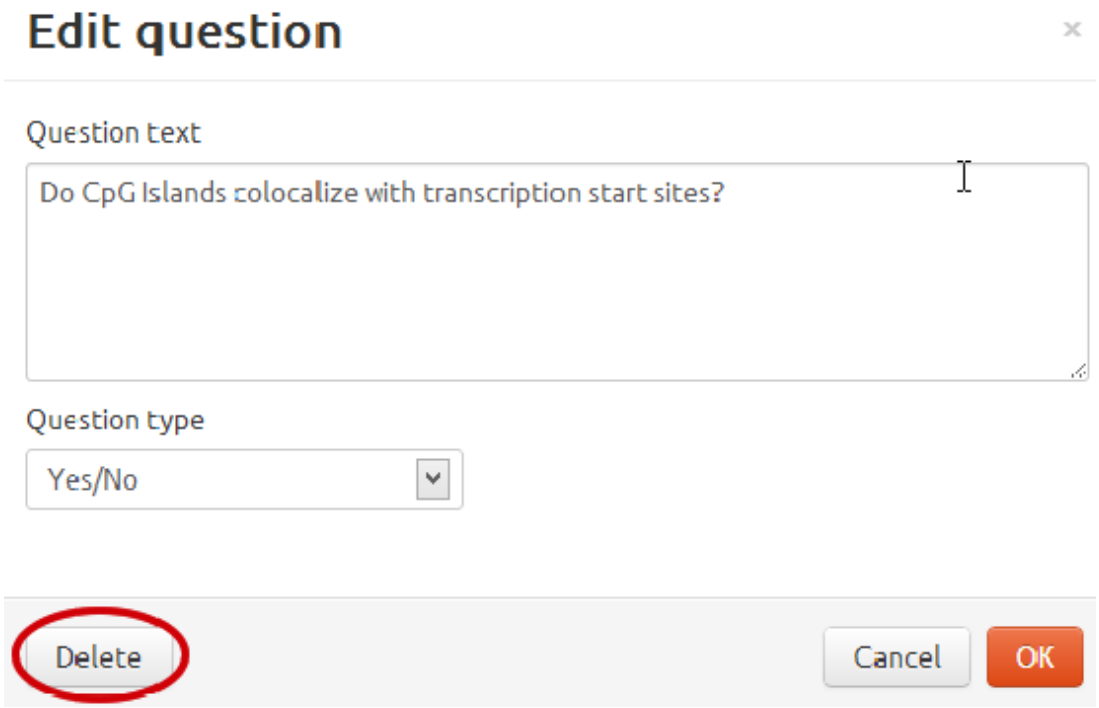


Figure 3.6: Edit/Delete question form.

Multi-word terms, like “di-glycine signature”, should be enclosed in quotation marks, as in the example above. The annotation tool attempts to retrieve concepts, articles, and statements that contain as many as possible of the specified terms. Recall that statements are entity relations retrieved from databases and knowledge bases, shown as pseudo-natural language sentences.

In Boolean queries, the terms are connected with AND and OR operators; brackets can also be used to clarify the scope of the operators.[1] Multi-word terms are again enclosed in quotation marks. For example, the following Boolean query retrieves articles that contain the term “disease” and (at the same time) at least one (or both) of the terms “quantitative trait loci” or “splicing”.

```
disease AND ("quantitative trait loci" OR "splicing")
```

Once the query has been entered, clicking on the “Search” button (Figure 3.7) executes the query.

## 3.4 Selection of concepts, articles, and statements

When the search specified by the query is completed, three lists containing concepts, articles (shown as “documents”), and statements appear (Figure 3.8). The contents of these lists can be viewed by clicking on the “Expand” links (Figure 3.8). The expert should select *all* the concepts that best characterize the question, *all* the possibly relevant articles (all the articles that the expert feels he/she should read or skim through more carefully), and *all* the statements that provide information that is useful to answer the question, as discussed in Step 4 of the guidelines of Chapter 2.

When a list is expanded, the expert can select items (concepts, documents, or statements) from the list by clicking on the corresponding “+” icons (Figures 3.9 and 3.10). When an item is selected, its “+” icon turns into a “–” icon. If an item has been accidentally selected, clicking on its “–” icon will remove it from the set of selected items. Figures 3.9 and 3.10 show examples of selecting concepts and documents respectively; the list of statements is very similar. In order to decide whether a document (article) is possibly relevant or not, the expert can view (inspect) it by clicking on its “i” icon (Figure 3.10). An “i” icon is also available for each concept and by clicking it some additional information concerning the concepts is displayed. Clicking on the page-like icon next to the “+” or “–” icon of an item displays the original source of the item (e.g., the corresponding PUBMED page for articles). Recall that the concepts come from biomedical terminology banks, databases, and knowledge bases (Chapter 2) and not all of them are appropriate for every query. For that reason, 5 buttons appear above the retrieved concepts (Figure 3.9). Each button corresponds to a resource from which concepts are retrieved. By clicking on these buttons, the expert can hide or show the retrieved concepts of the corresponding resources. An orange colour of the button indicates that the corresponding concepts are shown to the expert, while a grey colour indicates that they are hidden from the expert.

The items that have been selected for the question the expert is working with can also be be viewed in the drop-down box in the upper right corner of the annotation tool display (Figure 3.11).

## 3.5 Text snippet extraction

Having selected concepts, documents, and statements, the expert should now read (or skim through more carefully) the possibly relevant articles he/she selected. Clicking on the “Answer” tab of the upper navigation menu (Figure 3.12) shows all the items (concepts, documents, statements) that have been selected for the question the expert is working with (Figure 3.13). On the left of each item, a capital letter indicates the type of the item; i.e., “C” for concept, “D” for document, and “S” for statement

[1] Other operators are also available, but AND and OR should suffice in most cases.

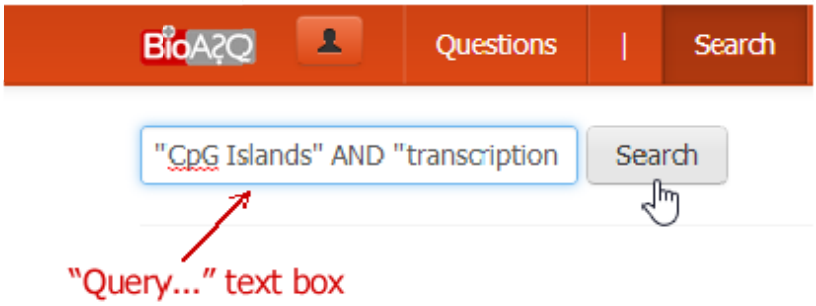


Figure 3.7: Performing a search.


Figure 3.8: Search results.

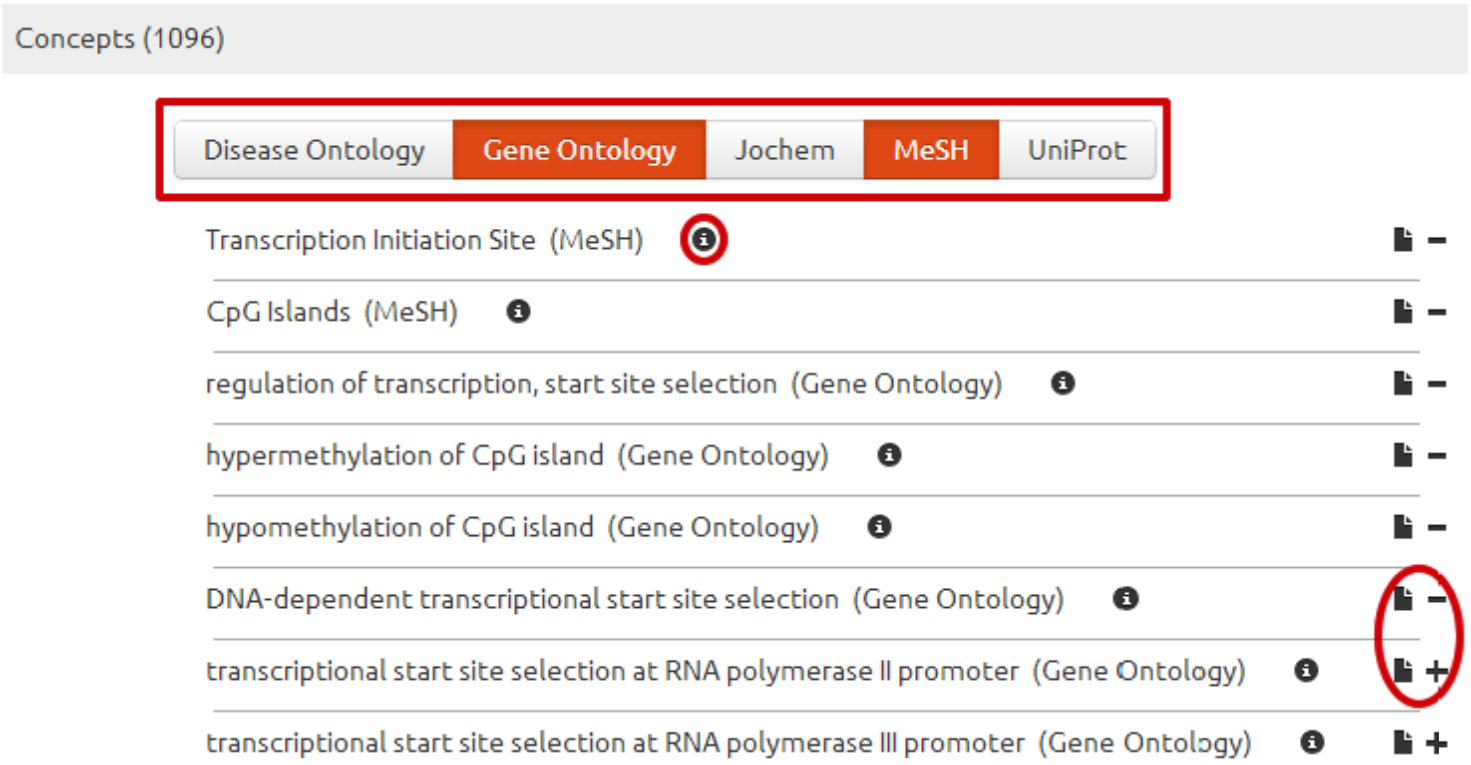


Figure 3.9: Concept selection.

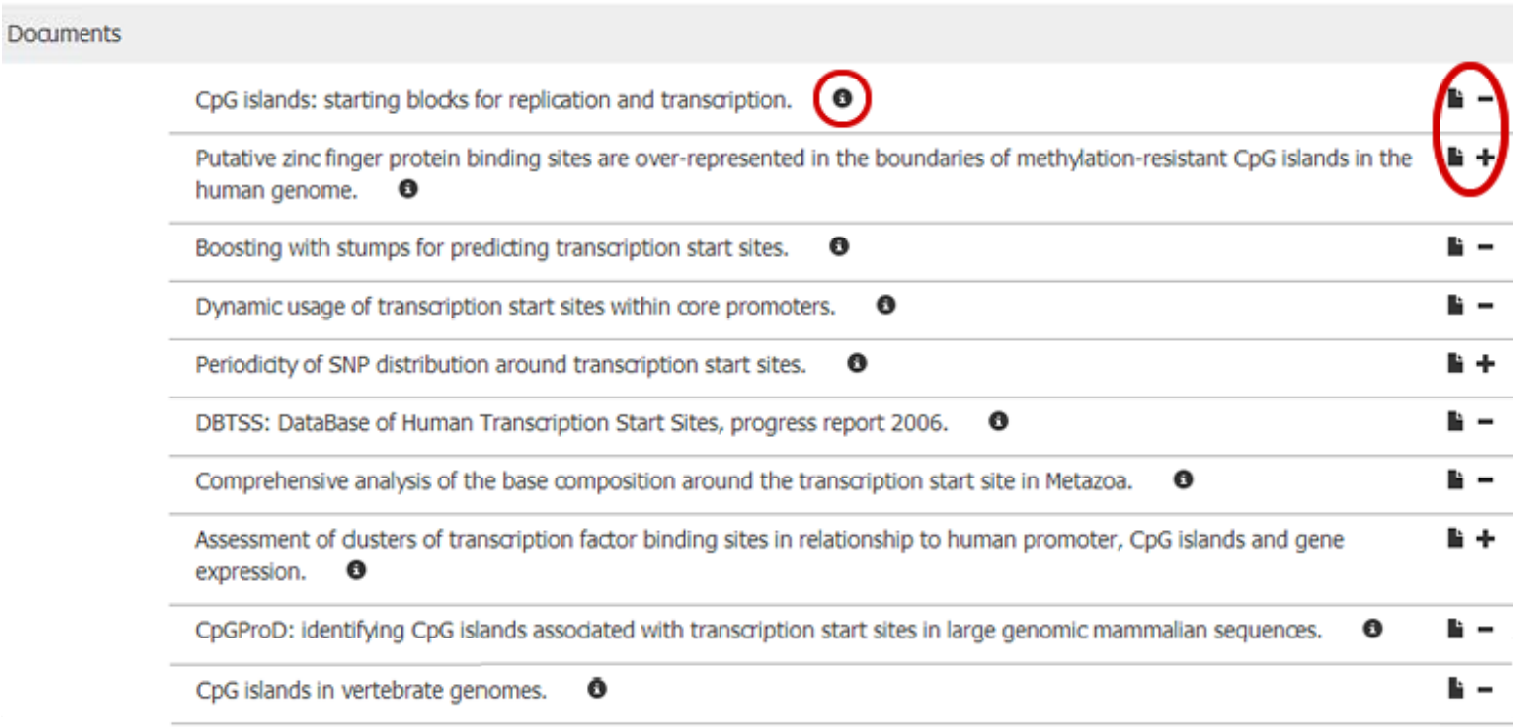


Figure 3.10: Document selection.

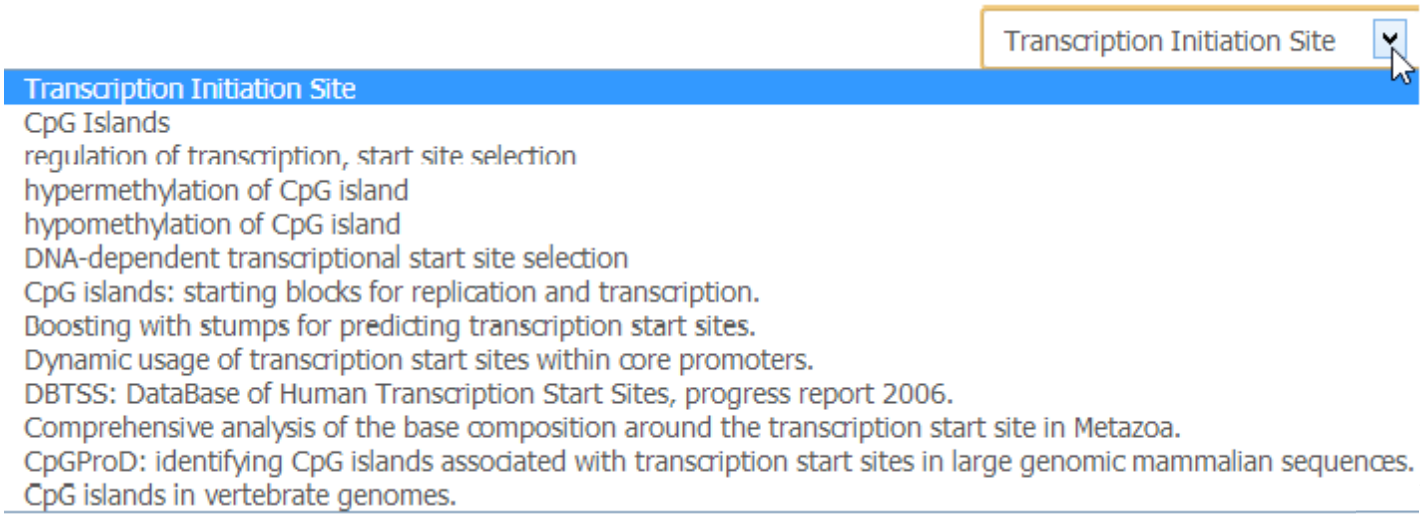


Figure 3.11: Selected items for a particular question.

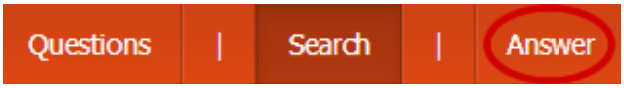


Figure 3.12: Selecting the "Answer" tab of the upper navigation menu.

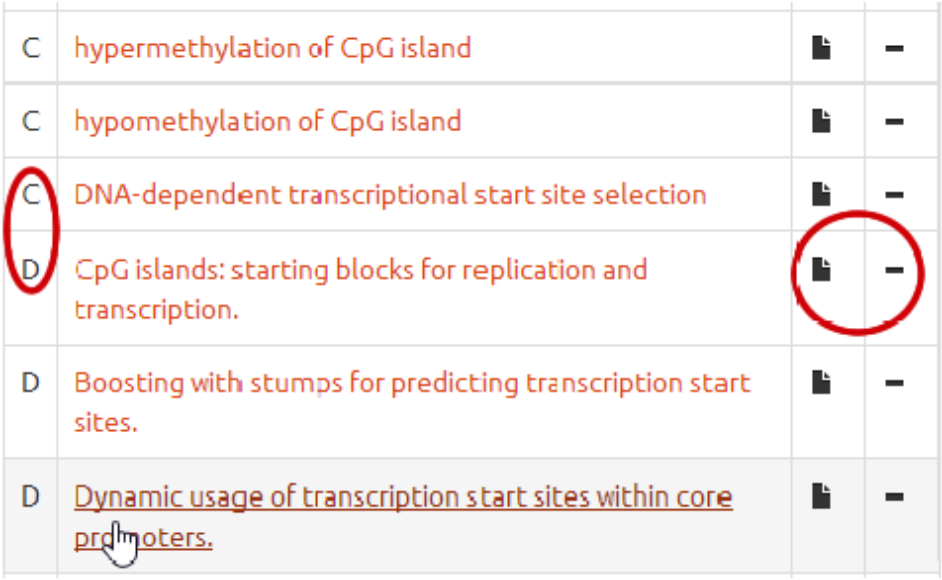


Figure 3.13: The selected items of a question, as shown when the "Answer" tab of the upper navigation menu is active. Only concepts and documents have been selected in this example.

(Figure 3.13. To remove an item (e.g., to remove a document that turned out not to be relevant), click on its "–" icon. Again, clicking on the page-like icon of an item displays the original source of the item (e.g., the corresponding PUBMED page for articles).

Clicking on the title of an article (document) displays the article (or its abstract, if only the abstract is available) and allows snippets to be extracted from the article (Figure 3.14), as discussed in Step 5 of the guidelines of Chapter 2. To extract a snippet, highlight it with the mouse and click on the "Annotate with selected snippet button" (Figure 3.14). The extracted snippet then appears highlighted in yellow. Clicking on the "X" button at the end of the snippet cancels the extraction (selection) of the corresponding snippet. At any time the expert can inspect the selected snippets by clicking on the "List of snippets link" right above the selected items (Figure 3.15). The expert may also delete a snippet from the list by clicking on the corresponding "X" button (Figure 3.15).

## 3.6 Query revision

If at this stage the expert feels that the selected statements and the extracted snippets do not provide enough information to answer the question, he/she should modify the search query, as discussed in Step 6 of the guidelines of Chapter 2. Clicking on the "Search" tab of the upper navigation menu (Figure 3.16) allows a new query to be entered, as discussed in Section 3.3. When the new query is executed, three new lists with the items (concepts, documents, and statements) retrieved by the new query appear (Figure 3.8), and the expert can again select the items that are appropriate. The items that had been retrieved by

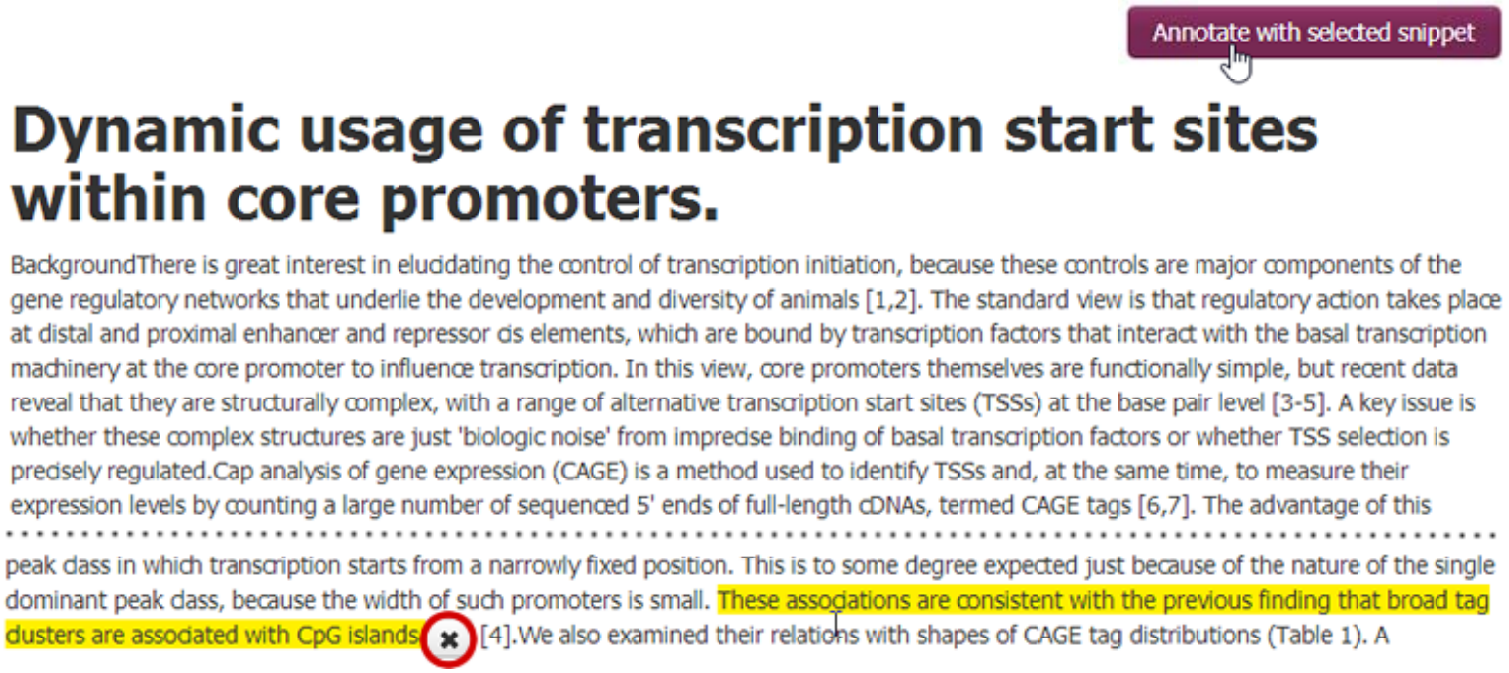


Figure 3.14: Extracting a snippet.

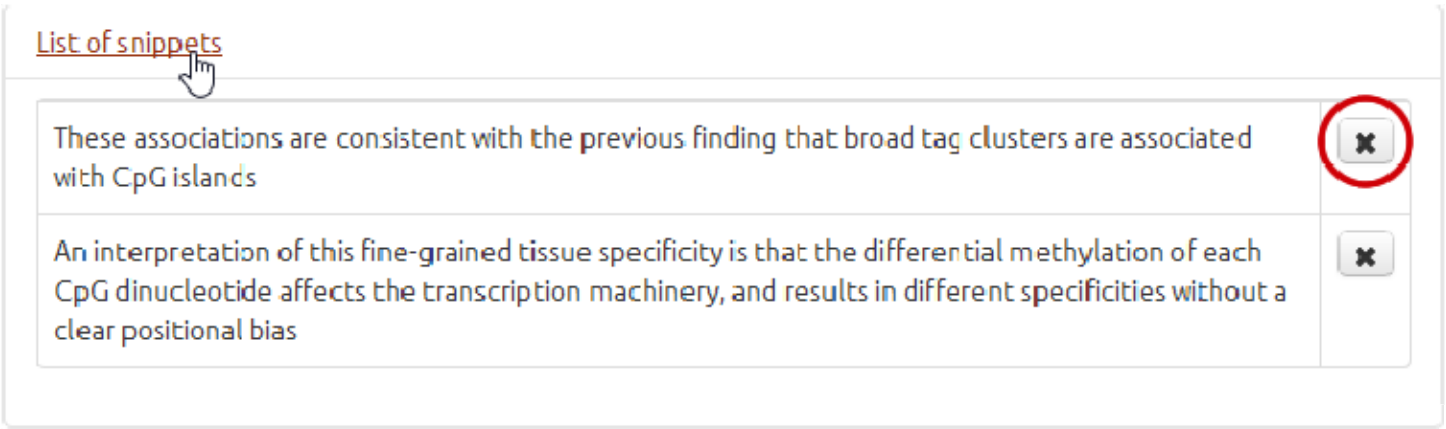


Figure 3.15: The list of selected snippets.

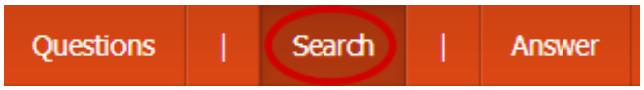


Figure 3.16: Selecting the "Search" tab of the upper navigation menu.

previous queries (for the same question) and had been selected by the expert (before executing the new query) are retained. All the items (from all the queries of the particular question) that have been selected are shown in the drop-down box in the upper right corner of the annotation tool display (Figure 3.11). They are also shown in the list of selected items (Figure 3.13) that appears when the "Answer" tab of the upper navigation menu (Figure 3.12) is active.

## 3.7 Exact and ideal answers

When the expert feels that the selected statements and the extracted snippets provide enough information to answer the question, he/she should formulate the exact answer and the ideal answer, as discussed in Steps 7 and 8 of the guidelines of Chapter 2.

Both the ideal answer and the exact answer (in that order) have to be entered in the text box (Figure 3.17) that appears when the "Answer" tab of the upper navigation menu is active (Figure 3.12). The text box will already contain a template text to be filled in. The ideal answer should be written immediately after the "Ideal answer:" line of the template text (Figure 3.17), and the "Ideal answer:" line should be maintained. The exact answer should be written immediately after the "Exact answer:" line of the template, and the "Exact answer:" line should be maintained. There should be an empty line between the last line of the ideal answer and the "Exact answer:" line.

For a yes/no question, the exact answer should be either "Yes" or "No" (Figure 3.17), with or without quotation marks; case does not matter (e.g., you may type "Yes", "yes", "YES" etc.). For a factoid question, the exact answer should be the name of the entity seeked by the question, enclosed in quotation marks, as in the following example; again case does not matter.

```
Exact answer:
"thalassemia"
```

If the entity has multiple names, all of them should be provided (to the extent possible), each one enclosed in quotation marks, with commas between the names, as in the following example:

```
Exact answer:
"influenza", "grippe"
```

For a list question, the exact answer should be the list of entities seeked by the question. The name of each entity should be written in a separate line, enclosed in quotation marks, with a double slash ("//") at the end of each line, as in the following example:

```
Exact answer:
"pneumonia" //
"bronchitis" //
```

If an entity (member of the list) has multiple names, all of them should be provided (to the extent possible) in the corresponding line, each one enclosed in quotation marks, with commas between the names of the same entity, and a double slash ("//") at the end of each line, as in the following example:

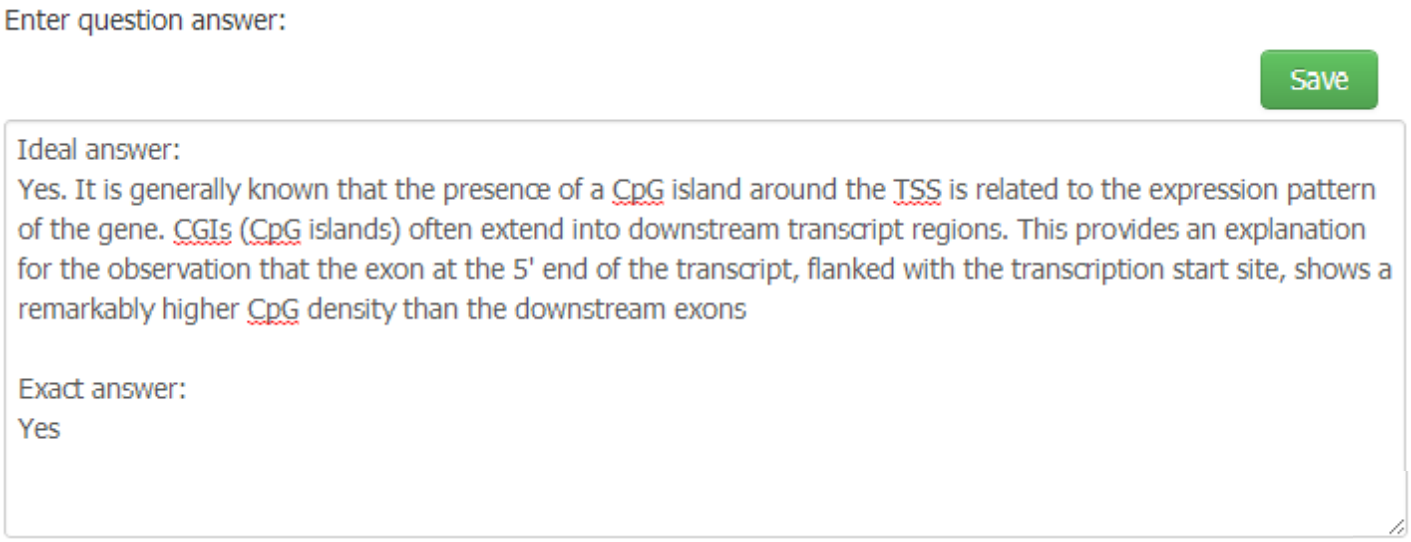


Figure 3.17: Entering the ideal and exact answer.

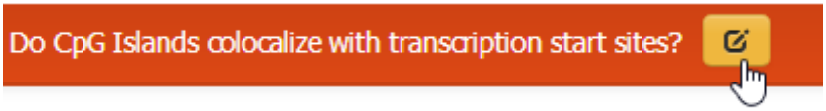


Figure 3.18: Editing the phrasing of the question.

```
Exact answer:
"pneumonia" //
"influenza", "grippe" //
"bronchitis" //
```

Clicking on the “Save” button (Figure 3.17) saves the ideal and exact answer that have been entered. A message will appear confirming that the ideal and exact answers have been saved.
**Important note: In some early versions of the annotation tool the “Save” button saves all the work that the expert has performed for a particular question, not just the ideal and exact answers. In these versions, all the work that has been performed since the last time the “Save” button was pressed remains unsaved, hence it is important to press the “Save” button often.**

## 3.8 Other useful functions of the annotation tool

The phrasing of the question the expert is working with can be changed at any time by clicking on the pencil-like button in the upper right hand corner of the annotation tool display (Figure 3.18). Once the phrasing of the question has been edited, it can be saved by clicking on the “✔” button (Figure 3.19).

To log out or to change password at any time, the person-like button of Figure 3.20 can be used. Clicking on that button leads to the form of Figure 3.21, where the expert can either log out or change his/her password.

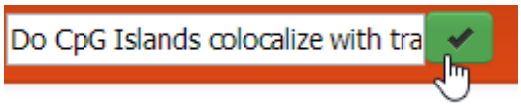


Figure 3.19: Saving the new phrasing of the question.


Figure 3.20: Logout or change password button.

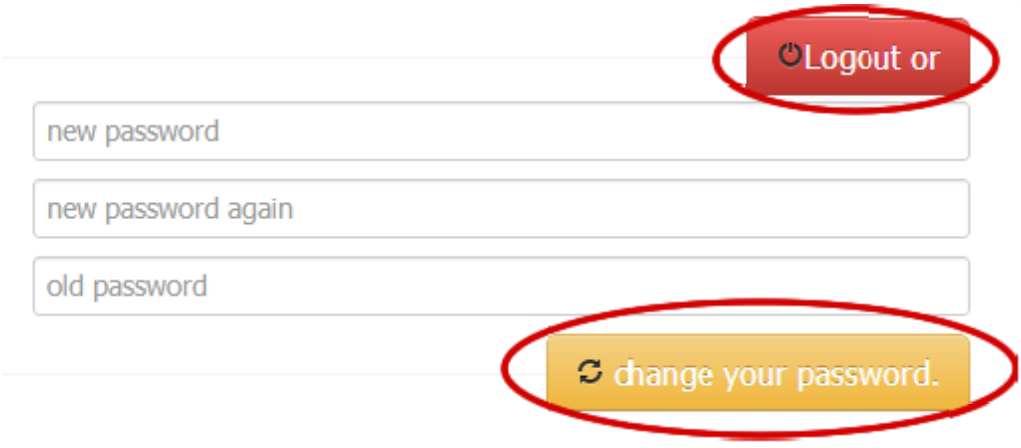


Figure 3.21: Logout or change password form.

# A

# Pilot study

A preliminary version of the guidelines was constructed by interacting closely with two biomedical experts, and by observing how they search for information during their research. The preliminary guidelines were then used in a pilot study, whereby each member of the biomedical expert team was asked to follow the guidelines in order to formulate one question, provide a reference answer, as well as information supporting the reference answer. Since the annotation tool had not yet been developed, the experts were asked to search for relevant articles only (not concepts and statements) using the search facilities of PUBMED and PUBMEDCENTRAL; the experts were also instructed to return their questions, reference answers, and supportive information in plain text form, along with any further feedback.

The preliminary guidelines that were used during the pilot study were the following:

**Step 1: Question formulation.** A question in natural language will be formulated. A question can be classified in one of the following categories:

**Factoid questions:** These are questions that require a particular entity as an answer; e.g., "What is currently the disease with the highest mortality rate in western countries?".

**Yes/No questions:** These are questions that require either "Yes" or "No" as an answer; e.g., "Are CNEs particularly enriched in gene deserts?".

**List questions:** These are questions that require a list of entities as an answer; e.g., "Which drugs are commonly used to treat HIV positive persons?".

**Other questions:** All other questions that do not belong in any of the previous categories; e.g., "What do you know about the H1N1 virus?".

The question should be as specialised as possible in order to contain the returned articles to a number between 10 and 20. This is a proposed range. More or fewer returned articles can also be accepted, provided that the volume of data is still manageable and that the set of retrieved relevant snippets is sufficient for answering the question.

**Step 2: Relevant terms extraction.** A set of relevant terms will be formed, which will include terms that are already present in the question, synonyms of the question's terms, closely related broader and narrower terms etc.

**Step 3: Article retrieval using PUBMEDCENTRAL (PMC).** The terms formed in Step 2 will be used to formulate a query (Boolean or simple bag of terms) which will be used to search PMC and a set of relevant articles will be retrieved. Recall that the number of articles should be around a range of 10 to 20. Feel free to use any "advanced search" provided by the PMC to enrich the results of the query.

**Step 4: Text snippet extraction and colouring.** From the set of articles retrieved during Step 3, extract *all* the text snippets containing information that can be used to answer the question of Step 1. This means that if there are text snippets that contain the same (or almost the same) information, all of them should be extracted and not just one of them. A text snippet is a piece of text containing useful information for the answer. Depending on the information each snippet contains, the snippets may be divided in two categories:

***Key* snippets:** These snippets contain information that is required in order to answer the question (i.e., the question cannot be answered without the information in these snippets).

***Supplementary* snippets:** These snippets provide extra information that is useful, but not required to answer the question.

The *key* snippets should be highlighted with red colour, while the *supplementary* ones with green. If all snippets retrieved are *key* snippets, they should be coloured red.

**Step 5: Query revision.** If the snippets extracted during Step 4 do not contain all the information needed to answer the question, the query should be revised with more or different relevant terms. The task will then continue from Step 3, i.e., the revised query will be used to search PMC and new text snippets will be extracted from the new articles retrieved. The revision of the query will continue until the expert feels that the extracted snippets provide enough information to answer the question.

**Step 6: Answer formulation and colouring.** Based on the text snippets of Step 4, create an ideal answer in natural language. The answer may be coloured in a similar way the snippets were coloured during Step 5. The parts of the answer providing key information (i.e., the parts that actually answer the question) should be highlighted with red color, while the parts of the answer with *supplementary* information (i.e., the parts without which the answer is still complete) should be highlighted with green colour. If it is unclear whether a part of the answer is important or supplementary, then this part should be coloured black.

**Step 7: Exact answer formulation.** For yes/no, factoid, and list questions, an exact answer should be formulated containing, respectively, "yes" or "no", the particular entity answering the question, or the list of entities answering the question.

After completing the 6 steps described above, the task should be repeated with the following modification: instead of using the articles of PMC, the abstracts of PUBMED should be used to extract the relevant snippets. Note that in this alternative query (using PUBMED) we proceed with a more extended database, but using only abstracts. Thus, it is very possible that the retrieval of a satisfactory set of relevant snippets and the formulation of a correct answer will fail, even after re-adjusting the used query terms. In such a case, this information will be included as a negative result in the final report. In fact we have observed such negative results with several example queries we have tested. Below you can find an example of the whole process using PMC.


Figure A.1: PMC search results for the Boolean query: (CNEs) AND (“gene deserts”)

**Step 1: Question formulation.** Are CNEs particularly enriched in gene deserts?

**Step 2: Relevant term extraction.** "CNEs", "gene deserts".

**Step 3: Article retrieval using PUBMEDCENTRAL (PMC).** We can formulate the following Boolean query: (CNEs) AND ("gene deserts"). Figure A.1 shows some of the 29 articles retrieved when searching PMC.

**Step 4: Text snippet extraction and colouring.** Below you can find the snippets extracted from the articles of Step 3.

- *"All but one of the CNE regions in human are located in gene-poor regions termed 'gene deserts' that flank or surround the trans-dev gene and are characteristic of regions thought to contain large numbers of cis-regulatory elements"*
- *"Here, we present a genome-wide survey of 10,402 constrained nonexonic elements in the human genome that have all been deposited by characterized mobile elements. These repeat instances have been under strong purifying selection since at least the boreoeutherian ancestor (100 Mya).They are most often located in gene deserts"*
- *"To further investigate the spatial congregation of exapted CNEs, we plotted the density of exaptations genome-wide, observing a very strong anti-correlation with gene density (Fig. 4). Indeed, the densest clusters are found in gene deserts"*
- *"Exaptation clusters are clearly most often found in large gene deserts"*
- *"Clustered CNEs are often found in gene deserts"*
- *"Studies using conserved non-coding elements (CNEs) and in-vivo GFP enhancer assays have shown that the majority of CNEs are located in gene deserts"*
- *"Five dCNE families were found to have no annotated paralogs in their vicinity. However, two of these families were located in gene deserts."*
- *"For dCNEs located in gene deserts, a search region up to the next known gene was used."*
- *"Nevertheless, the lack of alternative targets in these regions, as well as evidence that gene deserts harboring vertebrate-conserved elements are almost always adjacent to trans-dev genes (Ovcharenko et al. 2005), make it plausible that these elements and genes are indeed associated."*
- *"The dCNE on Chr5 is located within a 'gene desert' and is 9̃26 Kb3́ of the ISL1 translation start site."*
- *"In cases where paralogs were not identified and dCNEs were located in regions of low gene density (so-called 'gene deserts') we extended the region up to the next nearest gene"*
- *"Over 93% of the clusters (154/165) have a trans-dev gene located within 500 kb of one or more of its CNEs (Figure 2; Materials and Methods; Table S1). Of the remaining 11 clusters, five are closest to genes with zinc finger domains as identified by InterPro [46], one is in a gene desert, one maps to the AUTS2 gene region [47], and four are located adjacent to uncharacterised genes."*
- *"Five CNEs do not appear to cluster with any known genes in either the human or Fugu genomes and are located in a large gene desert on human Chromosome 22."*
- *"Interestingly, it has been shown that megabase deletions of two-gene deserts containing thousands of CNGs in mice had no phenotypic effects"*

**Step 5: Query revision.** In this example, no query revision is needed. However, a reasonable query revision, if one was needed, would be to use the term: "constrained non exonic elements". The revised Boolean query would be: ((CNEs) OR ("constrained non exonic elements")) AND ("gene deserts").

**Step 6: Answer formulation and colouring.** Yes, CNEs are most often found in gene-poor regions termed 'gene deserts'. There, they often form dense clusters.

**Step 7: Exact answer formulation.** Yes.

At the end of Stage 1, each biomedical expert should return two folders one for each search (i.e., one for PMC and one for PUBMED). Each folder should contain:

1. The question in natural language.
2. All the queries (in their Boolean forms, if Boolean queries were used), including the revised ones (if any). If "Advanced search" features were used they should also be reported with the respective queries.
3. The titles and URLs of the articles returned by PMC or PUBMED respectively, by all of the queries (original and revised ones).
4. For each returned article, all the relevant snippets coloured appropriately. If the query was revised any additional snippet should be included.
5. The answer in natural language, coloured appropriately if a colouring was needed.
6. The exact answer when appropriate.
7. Feedback concerning the task, i.e., anything they find useful to make the task better.

The biomedical experts carried out the pilot task successfully. Very few clarifications were required. The only part of the preliminary guidelines that seemed to cause confusion was the colouring of the snippets and ideal answers. There were a few cases, for example, where the same snippet was coloured both as red (key snippet) and green (supplementary). Additionally, the majority (74%) of the snippets where marked as key snippets. Furthermore, the requirement to colour the key and supplementary snippets and parts of the ideal answers turned out to be tedious and sometimes difficult; hence, it was removed from the guidelines of Chapter 2. The guidelines were also revised to make them clearer, taking into account the feedback of the pilot study, and they were then used as a basis for the design of the annotation tool. Once the annotation tool had been implemented, the guidelines were again updated, to take into account the functionality of the tool.