Querying and Managing Biological and Biomedical data Questionnaire

Thank you for participating in this study. All responses will be held in strict confidentially.

Organization:

Research interest:

This questionnaire is about the **data sources** you are used to query and the **tools** you are used to handle during your work. You can thus answer these questions by considering a particular context of your work (particular experiments' results you would like to better understand, etc.). Please describe below in a few sentences the context of the study you will follow in this document (what is your goal, what you are looking for…) and answer the following questions keeping this context of work in mind. You can also specify the context of your work more particularly during your answers.

Querying and Managing Biological and Biomedical data
Questionnaire

A. Data sources

1. In your work, do you query public data sources (SwissProt, GenBank…)? If yes, which ones (try to be as exhaustive as possible)?

2. Among them, which ones do you consult the most frequently?

3. How often do you consult these public sources?
   a-Everyday b-Twice a week c-Once a week d-Less

4. What kind of information do you search in these sources (annotations, sequences…)?

5. Choose a particular context from your own area of study and
   list some biological queries that you frequently make.
   For each query, (i) please check in the list below which entities are concerned (if some entities are missing, please specify), (ii) try to give synonyms of the verb(s) expressing the relationships between entities.

List of entities: BAC (=Clone), Chromosome, CytoBand, Domain, Disease, Experiment, Function, Gene, Marker, Pathway, Protein, Sample, SNP, STS, Patient.

Query	Entities concerned	Synonym(s) of the verbs
Example: *Return the proteins which 'map close' to domain GTARD…*	Protein, Domain	Map close = Similar with

6. How many data sources do you query when you look for a particular *entity?* For example, if you look for information about *proteins* do you only query SwissProt or also TrEMBL, PIR…? Give examples in your context.

7. If several sources yield answers for your query, do you access all of them or only few? If you query only a few, how do you proceed?

8. In your opinion, what is a "high-quality" source? (Please fill in this first side before turning over).

Concerning questions no. 9 and 10, **sort** the answers in order of importance in the context of your study.

9. When you look for information, what is the most important for you?
- Exhaustive answers (to collect all the results available)?
- Detailed answers (to collect few data but very detailed ones)?
- Reliable answers (to collect few data but validated ones)?
- No redundant answers?
- Easy-to-use answers (to collect data in a standard format)?
- Well-documented answers (to collect data with complete traceability of their origin)?
- Other (please specify):

10. In your opinion, a source of good quality is
- A source with many data (e.g. many sequences)?
- A source with detailed data (e.g. many features for each sequence)?
- A source with careful annotation (e.g. very high quality features for each sequence)?
- A source with no redundancy?
- A source with standard format?
- A source with up-to-date data?
- Other (please specify):

11. Could you think of sorting the answers you get according to some "quality criteria"? (Considering first answers provided by *sources with many data*, then sources *with reliable data*…)? If yes, please give examples of such criteria.

B. Cross-references, querying process

1. Are you used to follow cross-references between sources?
   Yes - No

2. When a source provides you with many distinct cross-references, how do you select the one(s) you follow? (Please do not read the following questions before answering this one).

Concerning question no. 3, **sort** the answers in order of importance in the context of your study.

3. Do you select the cross-references that you follow according to
   - The kind of information you are going to get?
   - The reliability of the source which is going to provide the data?
   - The fact that you know that this cross-reference has been manually added by an annotator and has not been automatically generated?
   - Other (please specify):

4. When you look for data related to two linked entities (e.g. a gene and the protein it encodes), how do you proceed (sources accessed, way of correlating information, etc.)?

(The following questions have been added from responses to the first interviews)

When you look for data about two *entities* linked together (e.g. a *gene* and a *protein* encoded by it):

5. Do you consider *neighbour entities* (e.g.*Disease*, *Pathways*...) which can be linked to *entities* of your query (gene and protein) or do you only consider data entities of your query?
For example, to answer the query already mentioned, are you interested in …

a. (with *neighbour entities*) … querying GenBank (to get data about *genes*), then OMIM (to get data about *diseases*), and finally SwissProt (to get data about *proteins*), following cross-reference links? In other words, you consider that your goal is to get information about *genes* and *proteins* and you may also get data about other *related-entities.*

b. (without *neighbour entities*) … querying only direct linked sources providing data about *genes* and *proteins*? In other words, you consider that taking into account other *entities* changes the meaning of your query.

Please illustrate your answer with an example of query given in section A.5.

6. Do you consider an order between the *entities* of your query?
For example, if you look for *proteins encoded by a given gene*, are you interested in…

1. (**ordered** entities) …querying sources which provide the given *gene* and then searching for associated *proteins*. In such case *entities* are ordered from *Gene* to *Protein*. For instance, you query GenBank and follow cross-references to SwissProt.
2. (**every order** between entities is considered) …considering entities in any order (from *Gene* to *Protein* and from *Protein* to *Gene*). In this case, you also take into account answers coming from SwissProt with cross-references to GenBank.

Please illustrate your answer with examples of query given in section A.5.

7. Do you only consider query involving *biologically linked entities?*
For example, if you look for information about a *disease* related to a given *3D-structure* of a protein, do you consider …

1. (**no restriction**) … any cross-references available? For example, you consider the cross-references between OMIM (providing *disease* information) and PDB (providing *3D structure* information) offering you information given by OMIM annotators. In this case, the cross-references do not correspond to a biological link because a disease has not a 3D structure.

2. (**restriction on biological links**) … only cross-references between sources providing *biologically linked* entities? For example, you consider cross-references from OMIM to SwissProt, then from SwissProt to PDB because *Disease* and *Protein* are biologically linked as well as *Protein* and *3-D structure.* You thus do not want to consider any information from two sources providing entities not directly biologically linked (you do thus not consider links from OMIM to PDB).

Please illustrate your answer with examples of query given in section A.5.

8. Do you search for information by considering one *entity* after the other?
(Entity-perspective). In other words, do you search for information entity by entity (first considering *Gene* then *Protein*…)?
Please illustrate your answer with examples of query given in section A.5.

9. Could you consider the same source many times (if this source gives you information about several *entities*)?
Please illustrate your answer.

10. Do you rather search for information by considering one source after the other?
(Source-perspective). Please illustrate your answer

C. Bioinformatics tools

1. In your work, do you use bioinformatics tools (e.g. Blast, …)?
   If yes, which ones?

2. Among them, which ones do you use the most frequently?

3. How often do you use these tools?
   a-Everyday b-Twice a week c-Once a week d-Less

4. How many tools do you use when you want to perform a given *bioinformatics task*?
   For example, if you look for *similarity between two sequences*, do you only perform a Blast or do you also use Fasta…? Please, give examples.

5. Are you used to change the parameters of a given tool (e.g. considering different substitution matrix in a Blast)? If yes, give examples and explain why.

6. Are you used to handle tools located
   a. Locally, in <u>your</u> workspace or in a computer of <u>your</u> organization?
   b. On the web, in big servers (e.g. NCBI, Expasy…)?
   c. Does it depend?

   If it depends on the tool, please specify the context.

7. In your opinion, what is a "high-quality" tool for you? (Please do not read the following question before answering this one).

8. What is the most important for you (sort the following answers)?
   A tool
   - standardized (e.g. Blast vs Fasta)?
   - which has been already used with many other experimental data and provides interesting results?
   - *Sure*? (nobody knows the way you use it and what your investigations are)
   - easy-to-use?
   - well-documented? (you want to know if the tool does exactly, what are the consequences of a given choice of parameters)
   - with an easy-to-read output?
   - Other (please specify):