Reflect – Augmented Browsing for the Life Scientist

Evangelos Pafilis[1], Seán I. O'Donoghue[1]*, Lars J. Jensen[1,2]*, Heiko Horn[1], Michael Kuhn[1], Nigel P. Brown[1], & Reinhard Schneider[1]*

[1]*European Molecular Biology Laboratory, 69117 Heidelberg, Germany.*
[2]*NNF Center for Protein Research, University of Copenhagen, Denmark.*
**These authors contributed equally.*

Correspondence to: Seán I. O'Donoghue e-mail: contact@reflect.ws

Anyone who regularly reads life science literature often comes across names of genes, proteins, or small molecules that they would like to know more about. To make this process easier, we have developed a new, free service called Reflect (http://reflect.ws) that can be installed as a plug-in to Firefox or Internet Explorer. Reflect tags gene, protein, and small molecule names in any web page, typically within a few seconds, and without affecting document layout. Clicking on a tagged gene or protein name opens a popup showing a concise summary that includes synonyms, database identifiers, sequence, domains, 3D structure, interaction partners, subcellular location, and related literature. Clicking on a tagged small molecule name opens a popup showing 2D structure and interaction partners. The popups also allow navigation to commonly used databases. In the future we plan to add further entity types to Reflect, including outside the life sciences.

As science uncovers the intricate interconnections within biological systems, many life scientists constantly come across unfamiliar biochemical entities (for example, genes, proteins, or small molecules) that were previously not known to be relevant to a given field, but where today's literature shows an important, new connection. For such cases, it is clearly valuable to systematically tag all scientific entities in a publication, thus helping the reader to navigate to more specific information about any entity of interest. Such tags can help the reader to comprehend scientific content more rapidly and completely. Even when an entity is already familiar to a reader, it can be valuable to have quick access to commonly used source data entries, for example protein sequences or 2D structures of small molecules.

In spite of the clear value of systematically tagging scientific entities, only a small fraction of the main scientific publishers currently offer such tags on their web content. Some publishers are beginning to explore the option of adding tags as part of the publication process[1], however enforcing, validating, and updating these tags creates additional work for publishers and authors.

The task of accurately tagging biochemical entities automatically is very challenging; this task has been the subject of intense research efforts that has lead to significant improvements in accuracy[2]. These automated methods have been used to develop a wide variety of text mining applications and services, many of which are designed to provide sophisticated search, analysis, and presentation capabilities[3]. However, a few text mining services have been designed to appeal to the broader life science community, for example iHOP[4] provides simple search, navigation, and presentation of Medline abstracts with systematically tagged gene and protein names.

Tagging a scientific entity is only half the story: the other half is the information that is accessed when the user clicks on a

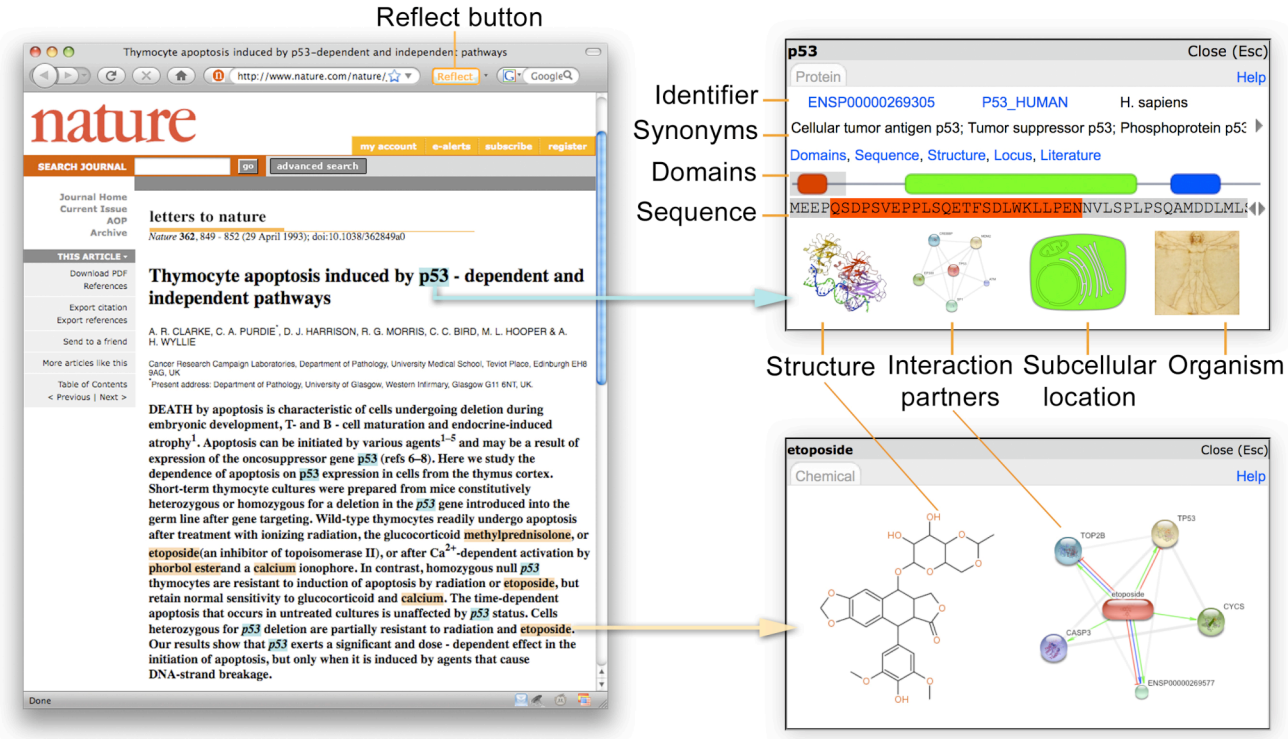

Figure 1 The Reflect button can be installed in Firefox or Internet Explorer. Clicking the Reflect button tags protein and gene names (blue highlighting), and small molecules (orange highlighting) in any web page. Clicking on a highlighted name opens a small popup showing a concise summary of important features of the entity, and provides access to related information (see METHODS for more details).

tag. In the past, entity tags were almost always simple hyperlinks to web pages showing source data entries. Increasingly, however, entity tags are not hyperlinks but scripts that create a small popup window (typically with Javascript). A key advantage of using popups is that users can see basic information about an entity without having to navigate away from the current web page. If needed, hyperlinks to more detailed information can be provided on the popup.

An emerging trend is to augment normal web browsing by using plug-ins such as Greasemonkey (http://greasespot.net) that let end-users modify the appearance of web pages while browsing. We believe that such augmented browsing tools will soon have an important impact on how scientists read literature on the web. For example, one such tool, ChemGM[5], lets end-users tag small molecule names in any web page; clicking on a tagged small molecule opens a popup that shows the 2D structure. Tagging is done by sending the page to a remote server, and the total time taken is typically about one minute for a five-page document. Another tool, Concept Web Linker (http://conceptweblinker.wikiprofessional.org), has a broader scope: it tags a range of entities, such as genes, chemicals, and diseases, again typically within about one minute. However, the Concept Web Linker popups show less specific information, giving only a short text description for each entity; to reach more specific information, such as protein sequences, the user needs to navigate through a series of web pages, in some cases browsing complex ontologies. A related system, Cohse[6], has even broader scope – it enables users to choose many different ontologies, including outside the life-sciences. Currently, however, the publicly accessible versions of Cohse provide only very limited functionality, and using the life-science ontologies provided does not allow direct navigation to specific information, such as sequences.

We designed Reflect to be an augmented browsing tool that would be broadly useful to life scientists, and would address the limitations of the above tools. A primary goal of Reflect was to enable the user to navigate directly from a gene or protein name to a specific sequence. A second goal was to be able to tag a typical web page in a few seconds. A third goal was to provide entity popups that give a concise summary of the most important features of the entities, as well as direct hyperlinks to commonly used source data entries (Fig. 1). Finally, Reflect was designed with a strong focus on ease of installation and on usability.

Reflect can be used directly from http://reflect.ws by typing or pasting in a URL. In this case, the Reflect server retrieves the HTML document, tags it, and returns the tagged version to the user's browser. Note that this will only work for URLs that are publicly accessible.

A more convenient way to use Reflect is to install it as a plug-in into Firefox or Internet Explorer. In this case, the HTML document is retrieved by the user's browser, then sent to the Reflect server, tagged, and returned to the browser. Thus, with the plug-in, users can 'Reflect' any page that they can access.

The Reflect server at the EMBL keeps in RAM (random-access memory) a large dictionary with names and synonyms for 4.3 million small molecules, and for 1.5 million proteins from 373 organisms. When tagging an HTML document, the server finds all occurrences of these synonyms, and returns a slightly modified version of the HTML document to the user's browser – the only difference is that all matching protein, gene, and small molecule names are now tagged and highlighted. Tagging a document usually takes much less time than uploading and downloading it; thus the time taken for the entire process (upload, tag, and download) depends almost exclusively on the speed of the user's internet connection. With standard broadband, the entire process usually takes from one to five seconds for a five-page document (See METHODS).

Clicking on a tagged small molecule name opens a summary popup (Fig. 2) that shows 2D structures from PubChem[7] and interaction partners from STITCH[8]. Clicking on a tagged protein or gene name opens a popup (Fig. 2) that shows synonyms, the complete amino acid sequence of the longest transcript, domains from the SMART[9] database, a representative 3D structure from PDBsum[10], principal interaction partners from STITCH[8], known sub-cellular location, and an image of the organism. Most of these features on the popup are hyperlinked to related database entries. The popup also has hyperlinks to the corresponding gene entry, and to related Medline abstracts in iHOP[4]. Dragging the mouse on the domain graphical view scrolls through the sequence, and hovering over a domain causes the domain name to appear in a tool-tip.

When a tagged name is ambiguous, the popup shows all possible matches and allows the user to disambiguate the name by choosing which of the possibilities is most appropriate. Currently, three levels of ambiguity are shown: first, a name may match both a protein and a small molecule; Reflect shows both possibilities on separate tabs. Secondly, a name may match to several genes within the same organism; Reflect shows all matching genes in a pull-down menu. Thirdly, for gene and protein names it is often ambiguous which organism is intended in the HTML document; Reflect shows a list of possible organisms, derived from the default organism (initially set to human, can be changed using the Firefox plug-in) plus organisms mentioned in the document. In the near future, we plan to show a fourth level of ambiguity, where users will be able to select splice variants for each gene.

Any automated method for recognizing biochemical entity names will make some errors: some false positive matches will arise due to overlap with commonly used words or acronyms, and false negatives will arise due to incompleteness of the tagging dictionary. To assess the accuracy of Reflect, we tested it against the BioCreative[11] benchmarks. Compared with 15 other tools for automated entity recognition that were assessed in BioCreative, Reflect ranked second best (91% F-score) using the Saccharomyces benchmark and had median performance (66% F-score) using the Drosophila benchmark. We consider these to be quite good results since, unlike the other tools tested against these benchmarks, Reflect was designed to optimize speed rather than accuracy.

In the near future we plan to enable community-based, collaborative editing for some of the information in Reflect popup, especially the synonym lists. These and other planned extensions will enable the user community to improve Reflect by correcting false negative and false positive matches. We plan to add further entity types (for example, diseases, pathways, and organisms), and eventually to add entity types beyond the life sciences; we designed Reflect to be an extendible platform, and we welcome collaboration proposals for adding further entity types. In addition, we welcome proposals from publishers and data providers interested in programmatic access to Reflect. With such access, end-users can use 'Reflected' content without needing to install a browser plug-in.

In summary, Reflect creates a view of the web tailored for the life scientist, that is, with systematic tagging of biochemical entities, and easy access to more detailed information. Reflect is already being used by thousands of researchers, and we have received much positive feedback regarding Reflect's usefulness and ease-of-use. In addition, just prior to this publication, Reflect was awarded first prize in the Elsevier Grand Challenge, a contest for tools that improve the way scientific information is communicated. Thus we believe that Reflect can be a valuable tool for researchers, teachers, students, and anyone who reads life science literature on the web. We further predict that in the near future tools such as Reflect will change dramatically how scientists use the web.

METHODS

User Interface. Both the browser plug-in and the web interface were constructed using HTML, JavaScript, XML-based User Interface Language, and Document Object Model events; communication between browser and server occurs via XMLHttpRequest objects.

Organism. By default, Reflect assumes protein names refer to human. The Firefox plug-in allows the user to change this default at any time to any of 373 organisms. In addition, the text of each HTML document is initially parsed to find recognized organism names, which are then added to the list of possible organisms for proteins in that document.

Tagging. After parsing for organism names, the text of the HTML document is parsed a second time for protein names. Parsing is done using leftmost longest matching of up to five words, testing each combination against the Reflect dictionary, which is stored in a hash table with all synonyms and orthographic variations occurring as hash keys. Recognized gene, protein, or small molecule names that occur in the text portion of the HMTL are then substituted with tags that call a Javascript function to generate the summary popups (Fig. 2). The document is then returned to the user's browser with previous HTML tags and their attributes unaffected, hence preserving the original document format.

Reflect Dictionary. The core component of Reflect is a consolidated dictionary that links synonyms to source data identifiers. The protein entries were derived from STRING[12], which in turn was created by importing the completed genomes in Ensembl, TAIR, Genome Review, and RefSeq (in this order of preference). In cases where one gene has several splice variants, the longest was chosen. In addition to importing all names and database accession numbers, we extended the dictionary with additional names from UniProtKB. The small molecule entries were derived from STITCH[8], which in turn was created by importing the compounds entries in PubChem[7]. Stereo-isomers were merged based on their canonical SMILES strings, and salt forms and trademark drug names were added as synonyms of the active substance. The dictionary is loaded into a Perl hash with each unique synonym

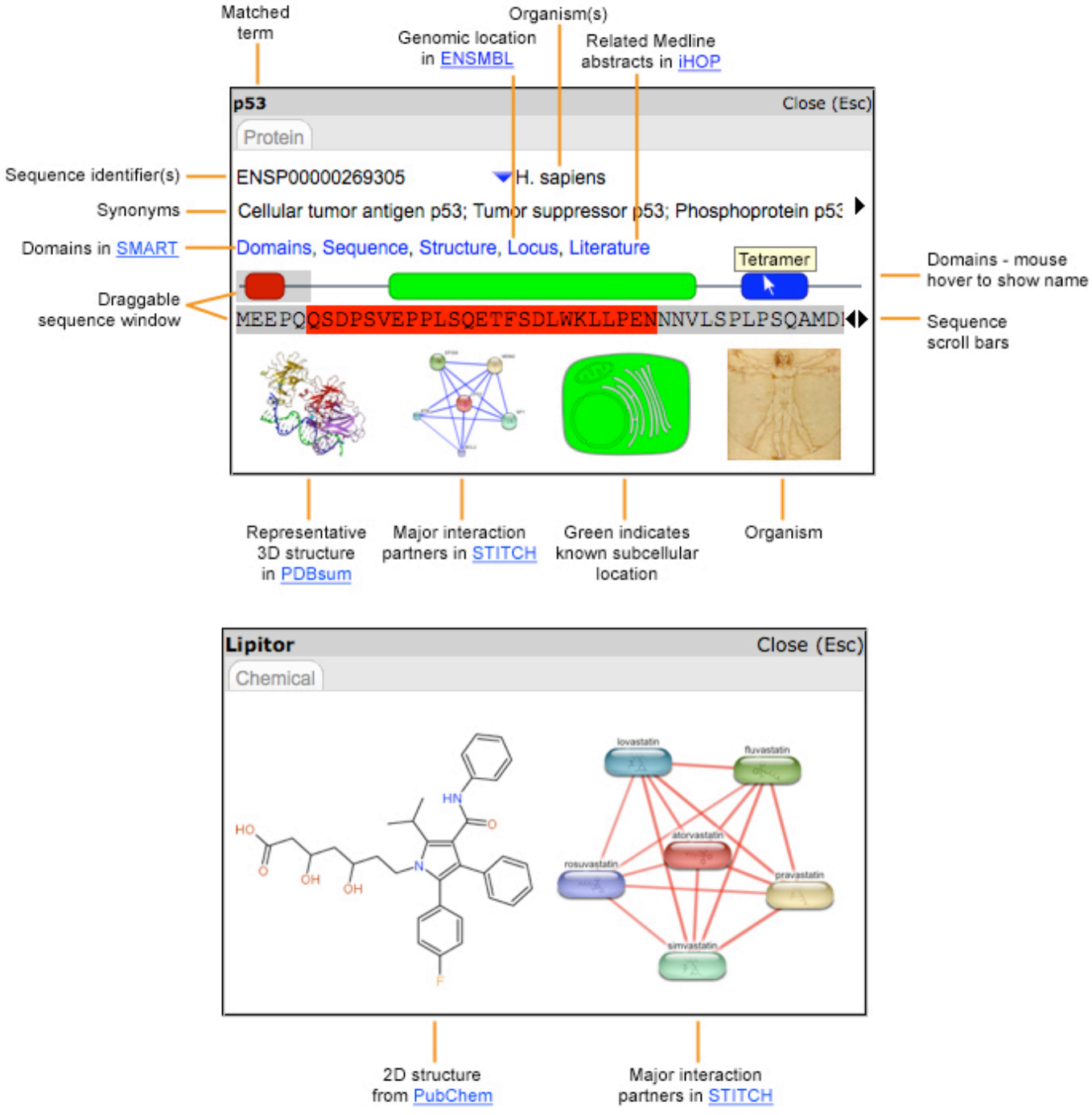

Figure 2 Details of summary popup content and links to more detailed information.

(including all orthographic variations) as a unique key; this enables fast tagging, currently at a cost of 18 GB of RAM.

Popups. The summary popups are generated using overLib (http://www.bosrup.com/web/overlib/) and with content supplied mostly by CGIs on the Reflect server. For proteins and genes, the popup shows the same synonyms used by the Reflect dictionary, except that database identifiers and orthographic variations are not shown. The Reflect database identifier is then used to fetch the sequence and domain information from SMART[9], and to fetch the image of the five most significant interaction partners from STITCH[8]. Scrolling of the sequence and synonym lists was implemented in Javascript. The links to the best matching 3D structure in PDBsum[10] and the information about subcellular location were pre-calculated from the sequence database entries. The organism images were taken from iTol[13]. For small molecules, the popup shows the 2D structure from PubChem[7], and the five most significant interactions are derived from STITCH.

Accuracy. The accuracy of Reflect tagging was assessed using the BioCreative[11] benchmarks for *Saccharomyces cerevisiae* and *Drosophila melanogaster* (task 1B). For both organisms, we used Reflect to tag 250 short texts, and we used the BioCreative 'gold standard' genes to calculate an F-score (equal to the geometric mean of precision and recall). BioCreative also includes a mouse benchmark, however we could not use it due to difficulties with converting the gene identifiers into those required for Reflect.

Extending Reflect. We designed Reflect to be an extendible platform that facilitates adding further entity types. For each entity type, the Reflect dictionary needs a list that maps each synonym to an identifier; in addition, for each entity type, Reflect needs the address of a web service that, when combined with an identifier, can create the popup content for a single entity.

ACKNOWLEDGEMENTS

Many thanks to Philippe Julien for the subcellular location viewer.

REFERENCES

1. Ceol, A., Chatr-Aryamontri, A., Licata, L. & Cesareni, G. *FEBS Lett* **582**, 1171-1177 (2008).
2. Smith, L. et al. *Genome Biol* **9 Suppl 2**, S2 (2008).
3. Krallinger, M., Valencia, A. & Hirschman, L. *Genome Biol* **9 Suppl 2**, S8 (2008).
4. Hoffmann, R. & Valencia, A. *Nat Genet* **36**, 664 (2004).
5. Willighagen, E.L. et al. *BMC Bioinformatics* **8**, 487 (2007).
6. Bechhofer, S.K., Stevens, R.D. & Lord, P.W. *Pac Symp Biocomput*, 79-90 (2005).
7. Wheeler, D.L. et al. *Nucleic Acids Res* **36**, D13-21 (2008).
8. Kuhn, M., von Mering, C., Campillos, M., Jensen, L.J. & Bork, P. *Nucleic Acids Res* **36**, D684-688 (2008).
9. Letunic, I. et al. *Nucleic Acids Res* **34**, D257-260 (2006).
10. Laskowski, R.A. *Nucleic Acids Res* **29**, 221-222 (2001).
11. Hirschman, L., Colosimo, M., Morgan, A. & Yeh, A. *BMC Bioinformatics* **6 Suppl 1**, S11 (2005).
12. von Mering, C. et al. *Nucleic Acids Res* **35**, D358-362 (2007).
13. Letunic, I. & Bork, P. *Bioinformatics* **23**, 127-128 (2007).