example, infusion reactions for 'lower risk' drugs can be clinically manageable and are usually of less concern than neutralizing antibodies cross-reacting with endogenous proteins. Nevertheless, ADAs for low-risk drugs can expose patients to clinical conditions that affect the overall benefit-risk assessment that is the central regulatory task before approval. The same holds true for a loss of efficacy by neutralizing antibodies. The situation is further complicated by the finding of different immune responses to a given product in different diseases and patient populations[4]. In other words, although overall a biological drug might be a lower-risk compound, the individual clinical sequelae are of high importance for the individual patient and have an impact on overall regulatory benefit-risk analysis.

Some explicit recommendations by Shankar *et al.*[1], such as the characterization of ADA positives to be explored only when considered necessary or decisions per clinical trial within a program, also pose the danger that data generated in a strategy for a predefined 'medically low risk' drug turn out to be insufficient for a proper benefit-risk assessment and thus do not meet regulatory requirements at the time of approval.

Thus, although many of the ideas developed by Shankar *et al.* have the merit of focusing on a risk-based approach and are valid and indeed interesting, we feel that their scheme of risk-based classification of products serve as a useful starting point for reflection at the time of planning immunogenicity studies as part of clinical trials; however, they may not in all cases be sufficient to support the benefit-risk evaluation required at the time of licensing of the product. Companies still need to justify their approach when filing a marketing authorization application in the European Union. We believe the European approach to immunogenicity, as presented in the final guideline document, retains a good balance by providing guidance on the conceptual planning of an immunogenicity evaluation on one hand, but being sufficiently open-minded to allow the flexibility needed for 'individual' biological drugs, on the other hand.

DISCLAIMER

C.K.S. is chairman of the CHMP Working Party on Similar Biological (Biosimilar) Medicinal Products Working Party (BMWP) and P.K. is a former chairman of BMWP. The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the EMEA or one of its committees or working parties.

Christian K Schneider[1,3], Marisa Papaluca[2] & Pekka Kurki[4]

[1]Paul-Ehrlich-Institut, Federal Agency for Sera and Vaccines, Langen, Germany. [2]European Medicines Agency, London, UK. [3]Twincore Centre for Experimental and Clinical Infection Research, Hannover, Germany. [4]National Agency for Medicines, Helsinki, Finland.
e-mail: schci@pei.de

1. Shankar, G., Pendley, C. & Stein, K.E. *Nat. Biotechnol.* **25**, 555–561 (2007).
2. European Medicines Agency. Guideline on Immunogenicity Assessment of Biotechnology-derived Therapeutic Proteins EMEA/CHMP/BMWP/14327/2006. <http://www.emea.europa.eu/htms/human/humanguidelines/multidiscipline.htm>
3. BMWP/BWP Workshop on Immunogenicity Assessment of Therapeutic Proteins, EMEA, London, September 4, 2007. <http://www.emea.europa.eu/meetings/conferences/4sep07.htm>
4. European Medicines Agency. European Public Assessment Report (EPAR) for Remicade. <http://www.emea.europa.eu/humandocs/Humans/EPAR/remicade/remicade.htm>

Reflect: augmented browsing for the life scientist

To the Editor:

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 web browsers, such as 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, three-dimensional structure, interaction partners, subcellular location and related literature. Clicking on a tagged small-molecule name opens a popup showing two-dimensional 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 those outside the life sciences.

As science uncovers the intricate interconnections within biological systems, many life scientists constantly come across unfamiliar biochemical entities (e.g., 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 two-dimensional 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 led 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 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 two-dimensional

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 those 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** and **Supplementary Methods** online). 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 work only for URLs that are publicly accessible.

A more convenient way to use Reflect is to install it as a plug-in to 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 European Molecular Biology Laboratory 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

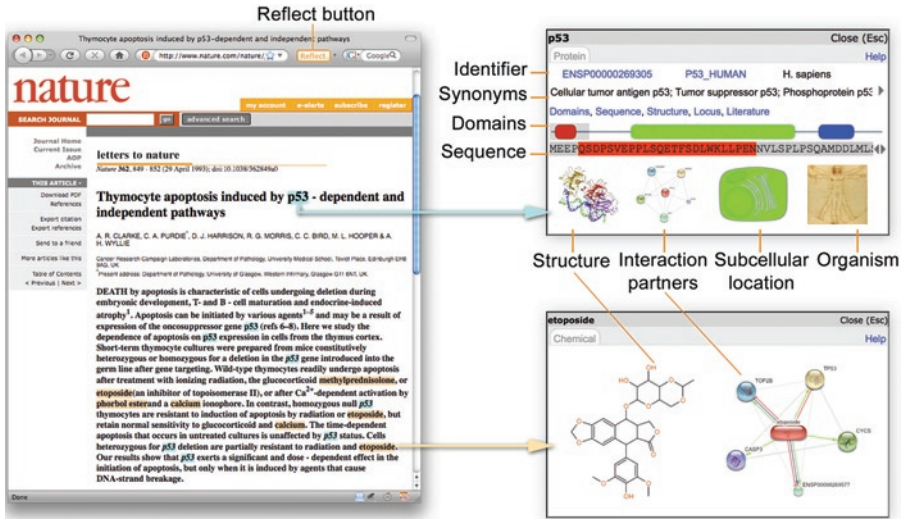

Figure 1 The Reflect button can be installed in the Firefox or Internet Explorer web browsers. Clicking the Reflect button tags protein and gene names (blue highlighting), and small molecules (orange highlighting) in any web page. Clicking on a highlighted name opens a small popup showing a concise summary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

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 (**Supplementary Methods**).

Clicking on a tagged small-molecule name opens a summary popup (**Fig. 1**, bottom right) that shows two-dimensional structures from PubChem[7] and interaction partners from STITCH[8]. Clicking on a tagged protein or gene name opens a popup (**Fig. 1**, top right) that shows synonyms, the complete amino acid sequence of the longest transcript, domains from the SMART[9] database, a representative three-dimensional structure from PDBsum[10], principal interaction partners from STITCH[8], known subcellular 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; in this case, Reflect shows both possibilities on separate tabs. Second, a name may match to several genes within the same organism; here, Reflect shows all matching genes in a pull-down menu. And third, for gene and protein names, it is often ambiguous which organism is intended in the HTML document; to address this, Reflect shows a list of possible organisms derived from the default organism (which is initially set to human, but can be changed using the Firefox plug-in) plus organisms mentioned in the document. In the near future, we also 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 cerevisiae* benchmark and had median performance (66% *F*-score) using the *Drosophila melanogaster* benchmark. We consider these to be quite good results because, 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 (e.g., 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 before publication of this correspondence, 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.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

Many thanks to Philippe Julien for the subcellular location viewer.

Evangelos Pafilis[1,3], Seán I O'Donoghue[1,3], Lars J Jensen[1–3], Heiko Horn[1], Michael Kuhn[1], Nigel P Brown[1] & Reinhard Schneider[1]

[1]European Molecular Biology Laboratory, Heidelberg, Germany. [2]NNF Center for Protein Research, University of Copenhagen, Denmark. [3]These authors contributed equally.
e-mail: contact@reflect.ws

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.* **10**, 79–90 (2005).
7. Wheeler, D.L. *et al. Nucleic Acids Res.* **36**, D13–D21 (2008).
8. Kuhn, M., von Mering, C., Campillos, M., Jensen, L.J. & Bork, P. *Nucleic Acids Res.* **36**, D684–D688 (2008).
9. Letunic, I. *et al. Nucleic Acids Res.* **34**, D257–D260 (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).

Extrapolating from sequence—the 2009 H1N1 'swine' influenza virus

To the Editor:

The recent incidence and spread in humans of the 'swine flu' influenza A virus has raised global concerns regarding its virulence and pandemic potential. The main cause of the so-called swine flu has been identified as human infection by influenza A viruses of a new H1N1 (hemagglutinin 1, neuraminidase 1) subtype, or '2009 H1N1 strain'. The first cases of human infection were reported in April in the Mexican town of La Gloria in Veracruz; soon after, reported infections occurred in areas of southern California and Texas. Several recent studies have focused on the necessary determinants for human adaptation and efficient human-to-human transmission of the H1N1 influenza A viruses[1–9]. Here, using a representative 2009 H1N1 strain as our starting point, we offer a perspective on the likely human adaptation and transmissibility of 2009 H1N1 viruses.

At the time when this sequence analysis was performed, partial or complete sequences were available from 38 different human isolates of the 2009 H1N1 virus. These sequences were obtained from GISAID (Global Initiative on Sharing Avian Influenza Data; http://platform.gisaid.org/) and the NCBI Influenza Virus Resource (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html). Comparison of the amino acid sequences between the 38 isolates showed some intragenic differences: seven amino acid positions in HA (hemagglutinin), one in M1 (matrix 1), two in M2 (matrix 2), four in NA (neuraminidase), three in NP (nucleoprotein), two in PA and two in PB2 (both of which encode subunits of viral RNA polymerase). Given the few intragenic variations among the 38 isolates available at the time of this study, we use /California/04/2009 (Cal0409) as a representative 2009 H1N1 virus strain for further analysis. The top ranking hits of the BLAST search using the individual Cal0409 genes are shown in **Supplementary Table 1** online.

Comparison of Cal0409 HA with the HA consensus sequences for human-adapted H1N1, avian-adapted H1N1 and swine-adapted H1N1 reveals important substitutions in positions 100–300, where the glycan receptor-binding sites and antigenic loops are located (**Supplementary Table 2** online). Notably, the Cal0409 HA possesses the signature amino acids Asp190 and Asp225 that have been shown to play a key role in conferring specificity to the human α2-6 sialylated glycan receptors[1]. We also observe amino acid substitutions that are unique to Cal0409 HA and have not been observed in previous human H1N1 HAs. These include substitutions at sequence positions 74, 131, 145, 208, 219, 261, 263, 264, 305, 317, 368, 377 and 530. Among these residue positions, 131 and 145 are proximal to the glycan-binding site.

To determine the possible effect of these mutations on the glycan-binding properties of HA, we constructed homology-based structural complexes of Cal0409 with representative α2-3 and α2-6 sialylated oligosaccharides, as described earlier[8] (**Fig. 1**). The construction of theoretical HA-glycan structural complexes previously[8] allowed us to provide a structural rationale for how specific amino acid mutations within the 1918 H1N1 HA can dramatically alter its relative α2-3/α2-6 binding affinity. Referencing these previous efforts, we determined the potential glycan binding properties of Cal0409 HA by analyzing its contacts with the α2-3 and α2-6 sialylated glycans.

On the basis of the observed contacts in the HA-glycan complexes, we summarize in **Table 1** the proposed roles of the residues in Cal0409 HA that provide binding specificity to α2-3 and α2-6 oligosaccharides, respectively. The main differences between the glycan-binding pockets of reference HAs and Cal0409 HA lie in the 140-loop region and the loop region preceding the 190-helix. Lys145, which is unique to Cal0409, along with Lys133 and Lys222, form a positively charged 'lysine fence' at the base of the binding site that potentially are positioned to anchor the N-acetylneuraminic acid (Neu5Ac) and galactose (Gal) sugars of both α2-3 and α2-6 glycans. In the case of the Cal0409 α2-6 oligosaccharide structural complex, the lysine fence also includes Lys156, which is positioned to provide additional contact with the glycan. The orientation of Asp190 is typically stabilized by a network of interactions involving residues at 186, 187 and 189 that precede the 190-helix. In Cal0409 HA, the residues at these positions are Ser186, Thr187 and Ala189; this set of residues is unique to the 2009 H1N1 strains. These residues appear to retain the ability to stabilize the orientation of Asp190 such that it is positioned to make optimal contacts with the third N-acetylglucosamine (GlcNAc) sugar (starting from Neu5Ac toward the reducing end) of α2-6 glycans, defined previously[8].

Our observations of the Cal0409 HA–glycan interactions suggest that this HA has the necessary residues to provide optimal contacts for