A gene network for navigating the literature

To the editor:

A network of genes and proteins extends through the scientific literature, touching on phenotypes, pathologies and gene function. We report the development of an information system that provides this network as a natural way of accessing the more than ten million abstracts in PubMed. By using genes and proteins as hyperlinks between sentences and abstracts, we convert the information in PubMed into one navigable resource and bring all the advantages of the internet to scientific literature investigation. Moreover, this literature network can be superimposed on experimental interaction data (*e.g.*, yeast-two hybrid data from *Drosophila melanogaster*[1] and *Caenorhabditis elegans*[2]) to make possible a simultaneous analysis of new and existing knowledge. The network, called Information Hyperlinked over Proteins (iHOP), contains half a million sentences and 30,000 different genes[3] from humans, mice, *D. melanogaster*, *C. elegans*, zebrafish, *Arabidopsis thaliana*, yeast and *Escherichia coli*.

Whereas conventional keyword and related article searches[4] result in long and not always informative lists of abstracts, navigation along the gene network allows for a stepwise and controlled exploration of the information space. Each step through the network produces information about one single gene and its interactions. Exploration of this gene-guided information network is intuitive and follows the associative organization of human memory[5], in which information is retrieved by connecting similar concepts[6]. The precision of gene name and synonym identification in iHOP ranges between 87% and 99% depending on the organism. Because researchers can move in iHOP between sentences taken directly from source abstracts, however, they always retain control over the reliability and relevance of information. This is an advantage over systems that translate the protein network from the literature into graphical representations[7], because these representations could give a misleading sense of confidence to the users and cloud the relevance of individual associations.

The iHOP system shows that distant medical and biological concepts can be related by surprisingly few intermediate genes; the shortest path between any two genes involves, on average, only four steps. We believe that this highly connected network will make human literature research more intuitive and efficient and also create a theoretical basis for the development of new automatic retrieval algorithms.

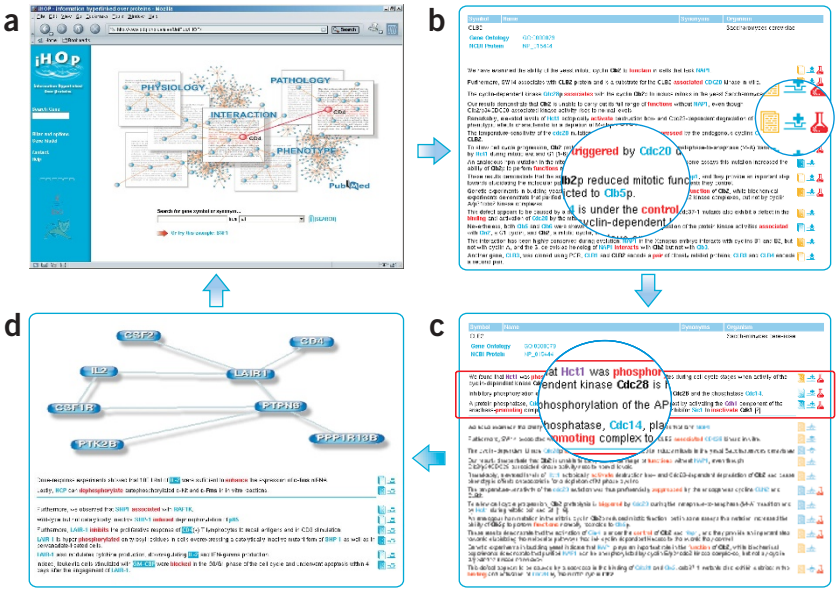

Figure 1 Navigation trail of iHOP. (**a**) The starting point for the literature investigation is a gene or protein of interest. (**b**) Information about a single gene X and its interactions is given as sentences taken directly from source abstracts. Gene names serve as hyperlinks to their corresponding pages in iHOP. Sentences that include proteins whose interaction has experimental evidence will be highlighted and ranked higher. All sentences are linked to the abstracts in which they appear, and abstracts from high-impact journals are also highlighted. (**c**) All sentences associating genes X and Y will be ranked first when the user arrives at gene Y from gene X. Thus, all information associating two genes is accessible on demand without obstructing the view of other associations in the network. Associating verbs are highlighted and influence the ranking positively[8]. (**d**) In the course of navigation through iHOP, interesting sentences can be collected into a logbook and are dynamically represented as a graph.

URL. The iHOP server is publicly accessible at http://www.pdg.cnb.uam.es/UniPub/iHOP/. Detailed descriptions of the text-mining methods and the technical architecture of iHOP will be published elsewhere.

ACKNOWLEDGMENTS

We thank the US National Library of Medicine for making MEDLINE publicly available and M. Tress and R. Allende for discussion. This work was supported in part by the ORIEL and TEMBLOR EC projects.

Robert Hoffmann & Alfonso Valencia

National Center of Biotechnology, CNB-CSIC, Cantoblanco Madrid M-28049, Spain. Correspondence should be addressed to R.H. (hoffmann@cnb.uam.es) or A.V. (valencia@cnb.uam.es).

1. Giot, L. *et al. Science* **302**, 1727–1736 (2003).
2. Li, S. *et al. Science* **303**, 540–543 (2004).
3. Hoffmann, R. & Valencia, A. *Trends Genet.* **19**, 79–81 (2003).
4. Kim, W., Aronson, A.R. & Wilbur, W.J. *Proc AMIA Symp.* 319–323 (2001).
5. Koch, C. & Laurent, G. *Science* **284**, 96–98 (1999).
6. Motter, A.E., de Moura, A.P., Lai, Y.C. & Dasgupta, P. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **65**, 065102 (2002).
7. Jenssen, T.K., Laegreid, A., Komorowski, J. & Hovig, E. *Nat. Genet.* **28**, 21–28 (2001).
8. Blaschke, C. & Valencia, A. *Genome Inform. Ser. Workshop Genome Inform.* **12**, 123–134 (2001).