BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii252–ii258
doi:10.1093/bioinformatics/bti1142

Text Mining

Implementing the iHOP concept for navigation of biomedical literature

Robert Hoffmann*,† and Alfonso Valencia

National Center of Biotechnology, CNB-CSIC. Campus de la UAM. Madrid E-28049, Spain

ABSTRACT

Motivation: The World Wide Web has profoundly changed the way in which we access information. Searching the internet is easy and fast, but more importantly, the interconnection of related contents makes it intuitive and closer to the associative organization of human memory. However, the information retrieval tools currently available to researchers in biology and medicine lag far behind the possibilities that the layman has come to expect from the internet.

Results: By using genes and proteins as hyperlinks between sentences and abstracts, the information in PubMed can be converted into one navigable resource. iHOP (Information Hyperlinked over Proteins) is an online service that provides this gene-guided network as a natural way of accessing millions of PubMed abstracts and brings all the advantages of the internet to scientific literature research. Navigating across interrelated sentences within this network is closer to human intuition than the use of conventional keyword searches and allows for stepwise and controlled acquisition of information. Moreover, this literature network can be superimposed upon experimental interaction data to facilitate the simultaneous analysis of novel and existing knowledge. The network presented in iHOP currently contains 5 million sentences and 40 000 genes from *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Danio rerio*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae* and *Escherichia coli*.

Availability: iHOP is freely accessible at http://www.pdg.cnb.uam.es/UniPub/iHOP/

Contact: hoffmann@cbio.mskcc.org

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The internet has fundamentally changed the way we access information; it is now a common practice to use search engines, such as Google™, and to follow hyperlinks rather than reach for a reference book (Schatz, 1997; Henzinger and Lawrence, 2004). In the internet, partitioning into pages and hyperlinking of related content makes retrieval of information associative and extremely effective. However, these revolutionary changes in the way we approach information are not reflected in the tools available for scientists.

Limits of automatic information retrieval. In recent years, the text-mining community has laid the foundations for the next generation of scientific information resources. Important advances have been made in the detection of biomedical entities within scientific text (Fukuda *et al*., 1998; Proux *et al*., 1998; Collier *et al*., 2000; Krauthammer *et al*., 2000; Friedman *et al*., 2001; Marcotte *et al*., 2001; Franzen *et al*., 2002; Hirschman *et al*., 2002; Yu *et al*., 2002; Hanisch *et al*., 2003; Morgan *et al*., 2003; Tsuruoka and Tsujii, 2003; Mika and Rost, 2004) and novel ideas for the analysis of large-scale experiments have also been introduced (Tanabe *et al*., 1999; Blaschke and Valencia, 2001; Friedman *et al*., 2001; Jenssen *et al*., 2001; Masys *et al*., 2001; Park *et al*., 2001; Raychaudhuri *et al*., 2002; Glenisson *et al*., 2004). However, from the very beginning of biomedical text-mining, the focus has been on automatic retrieval and extraction of information rather than on including the user dynamically in the discovery process. Current retrieval systems for the scientific literature are based on keyword searches and results usually take the form of long and not always informative lists of abstracts.

Interactive literature exploration. The World Wide Web, however, is so successful an information resource, because it grows naturally in view of a subsequent retrieval of information (Barabasi and Albert, 1999; Huberman and Adamic, 1999). Every time someone adds a link on a webpage, he/she tries to meet the interests of a potential visitor. Scientific publications contain also references to other publications; however, in the heterogeneous world of publishing houses and policies, it is difficult to make this reference network available for navigation; not all publications are electronically and freely available and a common standard for linking is not in sight. Moreover, the specificity of references varies strongly; references may point to a specific page or a book, to the discovery of one protein or the description of a complete protein network.

In practice, a topology has to be forced artificially upon scientific resources to make them navigable like the internet. The related article function, for example, attempts to bring order into PubMed by linking documents with similar contents (Wilbur and Yang, 1996). Although this method is very efficient, it can only provide links between complete documents, which are based on a mixture of concepts (e.g. terms, keywords, gene symbols) and which are thus less meaningful than links from one specific entity to a related document. The use of this approach for continuous navigation is, therefore, limited.

Biological information is naturally connected by genes. In other scientific fields, it will probably be very difficult to benefit from the advantages of hyperlinks in the exploration of textual information. In biology and medicine, however, we are in the exceptional position of having a natural underlying topology: in most biological models and hypotheses, genes and proteins serve as basic units of information. PubMed and any other resource of biological

*To whom correspondence should be addressed.

†Present address: Memorial Sloan-Kettering Cancer Center, MSKCC, New York, NY 10021, USA.

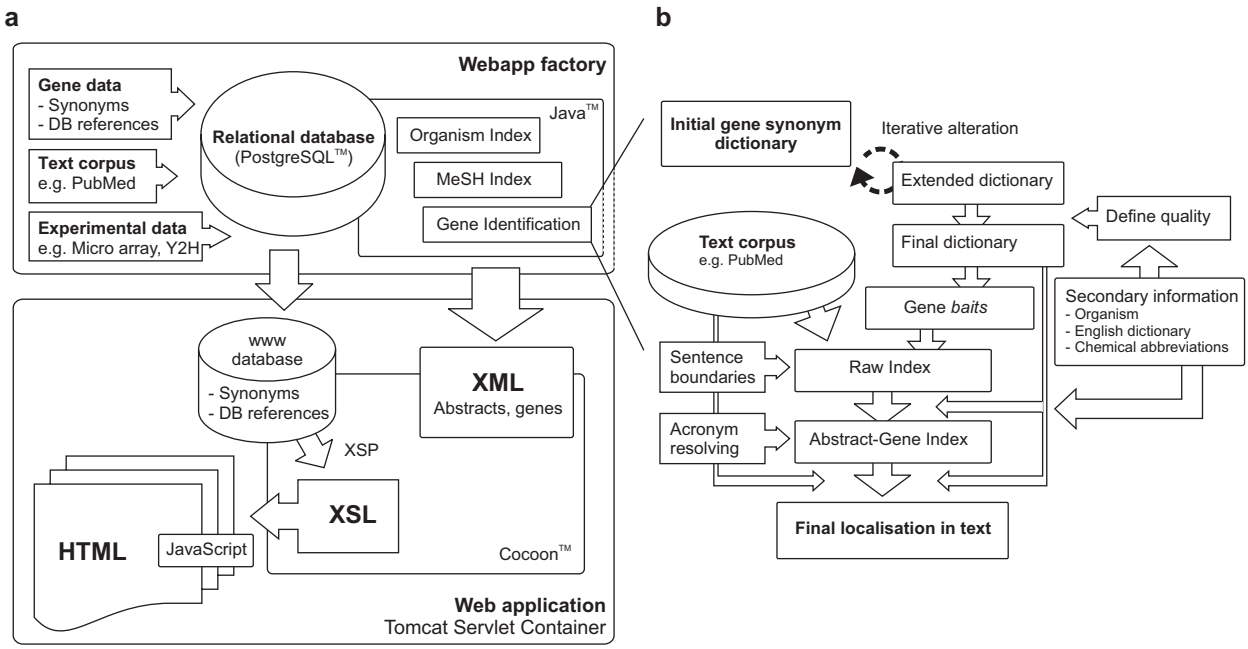

Fig. 1. (**a**) System architecture. The iHOP system is divided into two separate parts: the web application factory and the web application itself. Production state data in the web application is based entirely on XML technology and extremely fast response times are obtained through avoidance of complex front-end database queries. For every gene, one static XML document was created. Dynamic effects were achieved through the HTML and JavaScript layer on the client side to minimize server load. (**b**) Gene synonym identification in biomedical abstracts. More than 12 million abstracts were examined for the occurrence of gene symbols, names and synonyms. At an average abstract length of 200 words, the total number of examined terms reaches ∼2 billion, of which each could be one of the 3.2 million gene synonyms in the dictionary. To accomplish this comparison it was crucial to subdivide the gene identification process into independent steps of increasing precision, going from a raw gene–article index to a stable index, and finally to an exact localization of gene synonyms in the text.

knowledge can thus be seen as networks of concurrent genes and proteins that include relationships ranging from direct physical interaction (Ono *et al*., 2001) to less direct associations with phenotypic observations and pathologies (Jenssen *et al*., 2001; Perez-Iratxeta *et al*., 2002). In this paper, we utilize this network concept as a means of structuring and linking the biomedical literature and making it navigable in a way that is similar to the internet. In other words, we cluster and hyperlink the biomedical literature based on genes and proteins. The iHOP concept (Information Hyperlinked over Proteins) has recently been introduced to potential users (Hoffmann and Valencia, 2004); here, we discuss its feasibility and technical implementation.

2 SYSTEMS AND METHODS

The iHOP system consists of two separate parts: the iHOP factory and the iHOP web application (Fig. 1a). The factory manages and manipulates source data and produces the relevant XML output for the web application. The web application itself can be employed independently of the factory. Text and gene data are imported and organized within the factory in a relational database (PostgreSQL™). All software modules for import, data manipulation and export were developed in Java™.

External data are periodically updated (currently once a month) and consist mainly of gene synonyms and database references to external information sources (e.g. LocusLink or UniProt). The total number of source synonyms is about 530 000, and of database references 3 million (as of March 2005). Alongside the gene data, new abstracts are imported into the database and screened for organism and MeSH terms. The database schema also covers information on organisms (synonyms and NCBI taxonomy identifiers), a simple English dictionary and the complete MeSH thesaurus. The production of XML documents for abstracts and genes is divided into separate steps, going from importing gene data and updating the synonym dictionary to the actual gene detection. Computationally, the detection of gene synonyms within the text corpus is the most expensive step and was designed in such a way that it can be executed on a server cluster (queue system).

2.1 Gene synonym identification in biomedical text

Existing approaches for the identification of gene or protein synonyms in natural text are based on dictionaries, rules and machine-learning (Fukuda *et al*., 1998; Proux *et al*., 1998; Collier *et al*., 2000; Krauthammer *et al*., 2000; Jenssen *et al*., 2001; Morgan *et al*., 2003; Tsuruoka and Tsujii, 2003; Mika and Rost, 2004). The method implemented in iHOP is based on a dictionary approach to achieve maximum precision, and more importantly, to make cross-linking with external databases possible. Synonyms in the initial dictionary were derived from publicly available databases, e.g. LocusLink (Pruitt *et al*., 2000) and UniProt (Apweiler *et al*., 2004) and then iteratively extended to account for orthographical variations.

The actual indexing process was split into two main steps of increasing stringency (Fig. 1b). In the first step, abstracts were searched for parts of gene synonyms in a case-insensitive manner and by employing hashcode comparisons (Pieprzyk and Sadeghiyan, 1993) instead of character-by-character comparisons or regular expressions. In the final step, genes were assigned to precise positions in the text, taking all contextual information within abstracts into consideration (e.g. the organisms mentioned in or assigned to the text, explicit definitions, sentence boundaries, etc.). See the Supplementary Material for a detailed description of the synonym dictionary and the indexing process.

2.2 Web application

Starting from the final gene–article index, one XML document is produced for every gene and abstract. These documents essentially contain the original text divided into individual sentences with gene synonyms, MeSH terms and associated verbs tagged. Gene documents also contain general information, such as database references, synonyms, and a list of homologous genes, which in the web application are used to provide links to external resources.

Besides the precompiled XML data, which account for the largest part of the factory output, a small www-database is exported from the factory database. The main function of this database in the web application is to identify genes in the user query (by gene synonym or database reference) and to direct the user to the corresponding XML documents. The final navigable and hyperlinked HTML pages are created through transformation of the XML documents with XSL (Extensible Stylesheet Language; http://www.w3.org/Style/XSL/).

The web application thus consists only of the www-database, the XML-based data and the web archive (*war* file). The web archive includes all server side programs (XSP) and transformation style sheets for each type of XML document. Style sheet transformations (XSLT) are executed and cached in the Cocoon publishing framework (http://cocoon.apache.org/). Dynamic effects are achieved through the HTML and JavaScript layer on the client side to minimize server load and to avoid complex front-end database queries. This way, extremely fast response times are obtained and multiple concurrent usage of the system is possible. The war file can be employed under any Tomcat-compliant servlet container (http://jakarta.apache.org/tomcat/). See Supplementary Material for more information.

3 RESULTS

We developed a large-scale information system (http://www.pdg.cnb.uam.es/UniPub/iHOP) to demonstrate that it is scientifically beneficial and technically possible to convert large parts of the information in PubMed into one navigable information space. The results are divided into three sections that best demonstrate the expediency and feasibility of the iHOP concept. First, we analyse the singular potential of genes and proteins to divide the information in PubMed into meaningful and manageable clusters. Second, we describe precision and recall of the gene symbol identification, which is an essential part of the implementation of iHOP. Finally, we characterize the main properties of the emerging network.

3.1 Homogeneous division of medical and biological information

To convert PubMed into a navigable network it is necessary to find a classifier that divides the textual information into clusters of similar sizes and similar information content. The content of these clusters will be offered to the user as pages in the literature network and should not exceed a certain human manageable size.

The premise on which the current implementation of the iHOP concept was based is that genes and proteins are very specific and are able to divide the medical and biological literature into manageable sections of similar sizes. Indeed protein names are able to produce smaller clusters than other classifiers, i.e. MeSH terms.

MeSH is the thesaurus of the National Library of Medicine and was designed as a fast-access classification for PubMed (Kim *et al.*, 2001). We found that clusters based on MeSH terms generally contained more than 1000 articles, whereas 95% of all gene clusters were smaller than 100 abstracts (Fig. 2a). Even very specific MeSH terms (e.g. lymphomatoid granulomatosis) will retrieve an overly large amount of information from PubMed, whereas the search for a gene usually returns the information corresponding to a distinct molecular function or phenotypic effect.

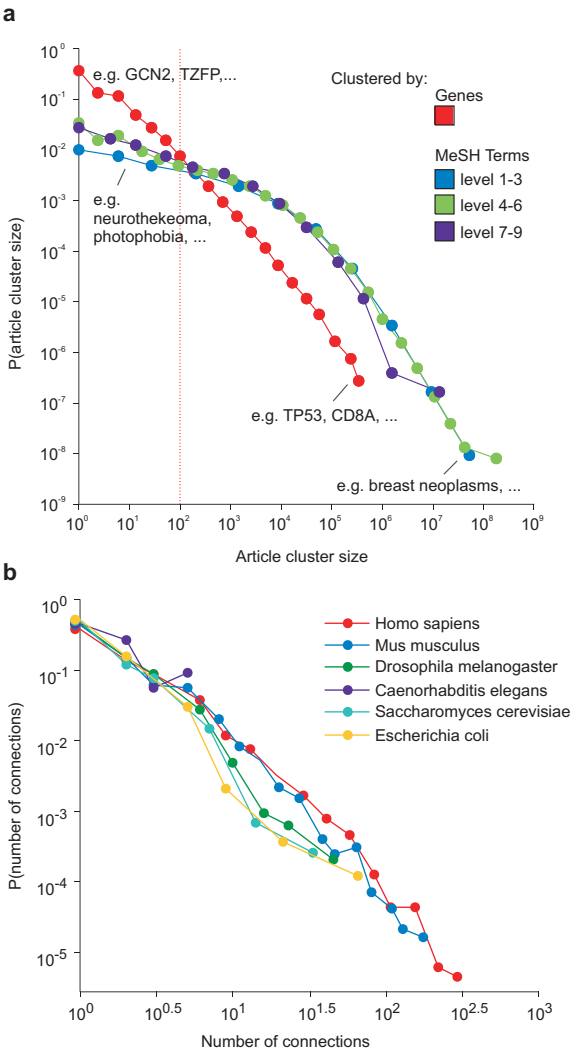

Fig. 2. (**a**) The distribution of document cluster sizes when clustering PubMed by MeSH terms and by genes. The figure shows that MeSH clusters generally contain more than 1000 documents, whereas 95% of all clusters grouped by genes are smaller than 100 abstracts. Genes and proteins are probably unique in that they divide PubMed homogeneously into clusters of manageable sizes that can be combined into one navigable network. (**b**) Connectivity distribution of genes in literature networks from well-studied organisms. The probability of identifying a given gene co-occurring with n other genes (in sentences of a gene–verb–gene pattern) is plotted on the y-axis and decays as a power law, appearing as a straight line on a log–log plot. A few central hub genes are associated with a large number of other genes, whereas the majority of genes only interact with a small number of others in the literature.

Furthermore, we tested whether clustering by genes and proteins also makes sense scientifically when compared with MeSH clusters. Clustering by universities, for example, might lead to small clusters, but does not reflect biologically relevant information. To assess the content of gene and MeSH clusters, we quantified the occurrence of specific terms (covering anatomy, diseases, physical and biological

Table 1. Global parameters of filtered gene networks extracted from the literature

	H.sapiens	*M.musculus*	*D.melanogaster*	*C.elegans*	*S.cerevisiae*	*E.coli*
Number of genes/proteins	4408	2723	696	193	854	534
Number of genes/proteins (largest connected subgraph)	4034	2374	562	25	534	410
Shortest path	3.4	5.1	5	4.4	6.42	5
Shortest path (random control)	5.2	5.7	7.3	5.7	10.6	7.8
Clustering coefficient	0.184	0.1514	0.156	0.1571	0.0892	0.1436
Clustering coefficient (random control)	0.0015	0.0013	0.0019	0.006	0.0004	0.001

science, chemicals and drugs) within each cluster and compared them with their overall background frequencies. We found that gene and MeSH clusters cover most domains to a comparable extent, with the expected exception that MeSH clusters show an emphasis on diseases, whereas gene clusters are focused more strongly on molecular aspects, e.g. genomic imprinting, regulation, oxidative stress (see Supplementary Material for details).

Overall, we conclude that genes and proteins fulfil the necessary requirements to divide PubMed homogeneously into meaningful clusters of a size that is more manageable for human experts than that produced by the MeSH thesaurus.

3.2 Recall and precision of gene identification

The automatic identification of gene and protein synonyms in natural language is an essential element of information extraction in medicine and biology (Fukuda *et al*., 1998; Proux *et al*., 1998; Tanabe *et al*., 1999; Pruitt *et al*., 2000; Friedman *et al*., 2001; Jenssen *et al*., 2001; Masys *et al*., 2001; Ono *et al*., 2001; Perez-Iratxeta *et al*., 2002; Stapley *et al*., 2002; Stuart *et al*., 2003). In the iHOP system, gene symbol and name identification is also an important component. In this fast-moving field, the modular design of iHOP will permit the exchange of gene-identification algorithms independent of the overall design and concept of the system.

However, gene identification was not a goal in itself, but rather, a necessary step in making the scientific literature navigable. We assessed recall of correct gene–article associations with reference data from the LocusLink database (Pruitt *et al*., 2000). Since some references can only be reproduced when considering the full text (Proux *et al*., 1998), only those gene–article references were evaluated in which at least one token of a gene synonym could be found case-insensitively within the abstract. Recall of gene–article references ranged between 65 and 91% depending on the organism. Recall in the manually curated BioCreative task1.B-corpus (Hirschman *et al*., 2005) ranged between 63 and 86%. Average recall of all documents about genes or proteins was 87%.

To estimate the precision with which gene synonyms were identified and precisely localized in the text, we manually evaluated about 400 gene text localizations per organism. Partially identified names or synonyms assigned to the wrong organism were considered false positives. We found that 3296 of all 3486 analysed text localizations were correct, corresponding to an average precision of 94%. The data used for the manual evaluation can be accessed at http://www.pdg.cnb.uam.es/UniPub/iHOP/info/gene_index/manual/index.html.

Harmonic F-measures, which combine precision and recall into comparable scores (Van Rijsbergen, 1979), ranged between 77 and 94%, depending on the organism. Currently, there are no other systems that cover all eight organisms or have been applied to the complete PubMed database; therefore, only partial comparisons are possible. However, most systems for more restricted domains publish lower or comparable F-measures (Fukuda *et al*., 1998; Collier *et al*., 2000; Franzen *et al*., 2002; Morgan *et al*., 2003; Tsuruoka and Tsujii, 2003; Mika and Rost, 2004). Exact results, problems and types of incorrectly identified genes are discussed in the Supplementary Material.

3.3 Protein networks in the literature are small worlds

All in all, 534 000 original synonyms and 3 million derivative synonyms for genes from *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Danio rerio*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae* and *Escherichia coli* were used to trace genes and proteins in 12 million PubMed abstracts (*New* PubMed). These identifiers were mapped back onto the original text, so that they could serve as hyperlinks between interrelated sentences. The network presented in the iHOP information system currently contains >40 000 different genes and ∼5 million sentences (as of March 2005). Of all sentences, 40% mention two or more genes and can be used for navigation. The remaining sentences mention a single gene together with one or more MeSH terms and provide comprehensive information on a gene's function, localization, etc. Table 1 compares the global properties of literature networks extracted for the six best-studied organisms.

The networks selected for this analysis were filtered for sentences that contain no more than two genes and that match a gene–verb–gene pattern. Verbs were chosen to cover the most frequent physical, enzymatic and regulatory aspects of interactions between genes and proteins. Of all verbs, 67% describe a regulatory relationship between genes, 26% discuss physical interactions and only 7% reflect enzymatic or other types of associations. A complete list of verbs and frequencies is provided in the Supplementary Material. These filtered networks are enriched in high-quality gene–gene associations compared with the complete but less specific network.

The networks of all analysed organisms were found to have small-world properties; thus, they exhibited short average paths between genes and elevated probabilities that neighbours of a given gene were also neighbours of each other (clustering coefficient). The average shortest path between any two genes involves between three and six steps, depending on the organism. Clustering coefficients of all analysed networks are about 100 times larger than of the random control networks.

The extremely short paths can be explained by the presence of literature hubs, genes that are discussed extensively in the context of others and that connect remote parts of the network. For instance, the shortest average path was found in literature on human genes, a fact that correlates clearly with the extremely high connected hubs in human research. The overall connectivity distribution is similar in all organisms and decays as a power law function (Fig. 2b); a few genes are highly connected in the literature, whereas the majority of genes have few interactions and are situated at the periphery of the network. The overall topology of these literature networks appears similar to those of experimental protein interaction networks (Jeong *et al*., 2001; Hoffmann and Valencia, 2003b); however, since the knowledge described in the literature is far from saturated, and sufficient experimental comparison data are missing, the true nature of this similarity remains to be determined. In particular, literature hubs (e.g. TNF, TP53 and IL2) have been shown not to be necessarily the most important genes in the cell (Hoffmann and Valencia, 2003a), although they play an important role in scientific research since they represent the best established facts and form a branching out point for the development of new understanding.

3.4 Navigating the small world

These small-world properties become manifest only when providing genes and proteins as hyperlinks. In iHOP this is realized such that every gene has one page containing all sentences that associate the gene with others. Gene synonyms within these sentences serve as hyperlinks to their corresponding pages. This way, each step through the network only produces the information pertaining to one single gene and its associations (Fig. 3a).

With the iHOP system it becomes evident that distant medical and biological concepts are related by surprisingly few intermediate genes. For example, the mouse T-box gene *Tbx3* has been shown to act as a repressor of cyclin-dependent kinase inhibitor 2A, thereby protecting against Myc-induced apoptosis, whereas mutations in human *TBX3* alter limb, apocrine and genital development in ulnar–mammary syndrome (Carlson *et al*., 2002; Davenport *et al*., 2003).

The basic concept of navigation was further enhanced by the weighting of sentences according to simple features and statistical parameters, such that the probability of finding relevant information first would be increased. Moreover, MeSH terms, associative verbs and high-impact journals were highlighted to facilitate the recognition of relevant information. Throughout, iHOP data are organized in XML, so that additional information can be included in layers, without having to change the basic framework or recompile the text corpus. At the current stage, experimental evidence for about 80 000 protein interactions from large-scale experiments for *D.melanogaster*, *S.cerevisiae*, *C.elegans* and the complete IntAct (Hermjakob *et al*., 2004) dataset are included to give additional confidence to sentences with specific gene–gene associations.

4 DISCUSSION

4.1 Advantages of textual networks over graphical representations

Currently, automatic information extraction methods can only support, but under no circumstances replace, human experts. The complexity of the information to be retrieved and the requirement for essential external knowledge make significant errors in all text-mining efforts unavoidable (Hirschman *et al*., 2002, 2005, http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html). For instance, there is no automatic system that solves the problem of synonyms colliding with symbols from other organism or common English words. Considering that about half a million new articles appear in PubMed every year, and that new genes are discovered and described continuously, gene name identification methods will probably always lag behind this creative process (Hoffmann and Valencia, 2003a). We, therefore, believe that approaches in which natural language is translated into logical or graphical representations are currently unrealistic and might have manipulative effects on the user's learning process, because these representations require a specificity that is not achieved by the algorithms employed, e.g. frames (Blaschke and Valencia, 2001), statistical methods (Jenssen *et al*., 2001), regular expressions, etc. Moreover, the graphical representation of large literature networks is, in practice, rather unsuited to the analysis and transmission of information. The sheer volume of information simply overtaxes most users, and more importantly, the accuracy of the extracted information varies significantly across the network (or any other abstract representation). In practice, this means that the user is forced to check many connections manually, which involves switching back and forth between the text source and the graph. Thus, the creative process of gathering new information and the generation of hypotheses is far from intuitive and is often frustrating.

The concept of iHOP is that researchers can move between sentences taken directly from their source abstracts and thus retain control over the reliability of the information they obtain. In this way, the expert will be in a position to recognize incorrectly identified symbols, in many cases simply by scanning a sentence, and more importantly, the expert will recognize whether the information offered in a sentence is of any interest at all. Both are tasks that no system can perform automatically with a comparable precision or without significant loss of recall. An example that illustrates this problem is the assignment of a gene synonym to the correct organism, particularly in cases in which the same synonym is used in two organisms and only differs in the composition of upper and lower cases. In fact, a large proportion of false negatives (34%) can be explained by synonyms that were correctly identified but then assigned to the wrong organism. Although nomenclature guidelines (i.e. HUGO) indicate that human gene symbols should be preferentially in upper case and homologous mouse genes in lower case, authors do not always observe these guidelines, especially when the homology of two genes forms a part of their argument (e.g. mouse 'Mtx2' and human 'MTX2'). In iHOP, contextual information is used to resolve these ambiguities automatically when possible. If, however, this is not possible, the user is provided a multi-link to all possible options, so that continuous navigation can be assured. This way, the user has the final say in interpreting the information provided and determining its relevance.

Consequently, iHOP was not designed to replace the expert, but rather to lead him/her as directly as possible towards the required information. In the future, we expect similar systems to be trained by the behaviour of their users (e.g. user decisions on multi-links).

4.2 Continuous navigation trail

The concept of a navigable scientific literature resource depends in practice on a continuous and intuitive navigation trail. In iHOP, we mimic standard internet navigation to maintain a flat learning curve. For example, the user might be analysing sentences for gene A and then come across an interesting or unexpected association with

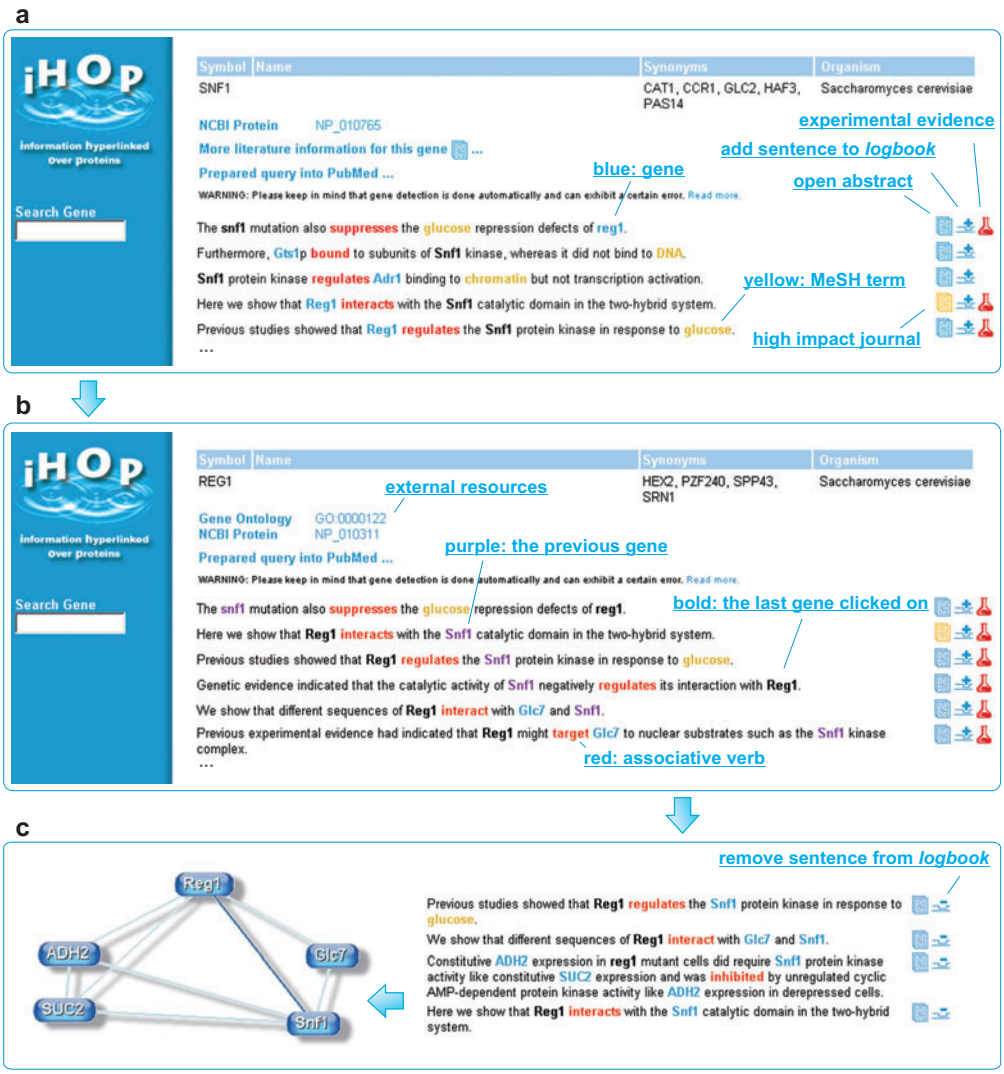

Fig. 3. Navigation trail of iHOP (Information Hyperlinked over Proteins). (**a**) The starting point for the literature investigation is a gene or protein of interest. Information pertaining to a single gene (e.g. SNF1) and its interactions is provided in the form of sentences taken directly from their source abstracts. Gene names serve as hyperlinks to their corresponding pages in iHOP. Sentences that include proteins whose interaction is experimentally supported will be highlighted and ranked higher. All sentences are linked to their abstracts and abstracts from high-impact journals are also highlighted. (**b**) All sentences associating A and B (here SNF1 and REG1) will be ranked first when the user arrives at gene B from gene A. In this way, all information associating two genes is accessible on demand without obstructing the view on all the other associations within the network. (**c**) In the course of navigation through iHOP, interesting sentences can be collected into a logbook or gene model and are dynamically represented as a graph.

gene B. The sentence found will mention one aspect of the relationship between A and B, but not necessarily the most important one. By clicking on the synonym of gene B, however, its corresponding information page will be shown, and all sentences associating A and B will be ranked first, since the user arrived at gene B from gene A. In this way, all information associating two genes is accessible on demand without obstructing the view on all the other associations within the network (Fig. 3b). We expect the exploration of this emerging network to be intuitive, since it is known that human memory is associative and that information is retrieved by connecting similar concepts (Koch and Laurent, 1999; Motter *et al.*, 2002).

Another important feature to make navigation efficient is a history function, something that allows users to keep track of their movements through the network. In the course of navigation through iHOP, interesting sentences can be collected into a logbook or gene model. The concept of the logbook was developed to collect newly acquired information and to generate consistent interaction models. The logbook also includes a dynamic graph to provide a graphical representation of all associations between the genes in collected sentences (Fig. 3c). This graph represents the condensed result of a literature search, but also remains hyperlinked to the corresponding sentences. In this way, users can familiarize themselves with the newly acquired information in an interactive manner and further extend the model. In contrast to automatically extracted graphs, this graph was created directly by the user and is therefore expected to fulfil the user's own expectations concerning amount and quality.

4.3 Superimposition of the textual network with experimental data

Recently, large-scale experiments have provided novel insights into intracellular interaction networks (Uetz *et al*., 2000; Ito *et al*., 2001; Gavin *et al*., 2002; Ho *et al*., 2002; Giot *et al*., 2003). However, this increase in information comes at the cost of reduced quality compared with conventional experiments (Bader and Hogue, 2002; von Mering *et al*., 2002; Hoffmann and Valencia, 2003b) and manual evaluation by experts remains indispensable. Manual evaluation usually includes comparison with existing information from the literature, and as such, is an expensive and time consuming process. We believe that the iHOP concept is naturally suited to combining the analysis of novel experimental and existing knowledge, facilitating, for example, the construction of reviews around protein networks. In iHOP, experimental data can be overlaid upon the textual network such that the sentences that include proteins whose interaction is experimentally supported will be highlighted and weighted more strongly (Fig. 3a).

To date we have implemented publicly available data from various large-scale experiments to demonstrate the strength of this approach. However, this approach is not limited to protein networks, since many datasets can be represented as networks. Microarray expression data, for instance, can be transformed into networks, where edges would correspond to gene pairs that exhibit highly correlated expression profiles (Stuart *et al*., 2003). In the future we will provide the user with the possibility of uploading additional experimental data for individual analysis in iHOP. Moreover, we plan to extend the iHOP concept to other text sources, especially to figure legends and full text articles from open access journals.

Overall, we believe that the iHOP concept will make human literature exploration more intuitive and efficient, while serving as a theoretical basis for the development of novel automatic retrieval algorithms in medicine and biology.

ACKNOWLEDGEMENTS

We also thank Michael Tress for helpful discussion. We are grateful to the US National Library of Medicine for making MEDLINE publicly available. This work was supported in part by the ORIEL (IST-2001-32688) and TEMBLOR (QLRT-2001-00015) EC projects.

Conflict of Interest: none declared.

REFERENCES

Apweiler,R. *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32** (Database issue), D115–D119.

Bader,G.D. and Hogue,C.W. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991–997.

Blaschke,C. and Valencia,A. (2001) The potential use of SUISEKI as a protein interaction discovery tool. *Genome Inform. Ser. Workshop Genome Inform.*, **12**, 123–134.

Carlson,H. *et al*. (2002) Tbx3 impinges on the p53 pathway to suppress apoptosis, facilitate cell transformation and block myogenic differentiation. *Oncogene*, **21**, 3827–3835.

Collier,N. *et al*. (2000) Extracting the names of genes and gene products with a Hidden Markov Model. *Proc COLING 2000*, 201–207.

Davenport,T.G. *et al*. (2003) Mammary gland, limb and yolk sac defects in mice lacking *Tbx3*, the gene mutated in human ulnar mammary syndrome. *Development*, **130**, 2263–2273.

Franzen,K. *et al*. (2002) Protein names and how to find them. *Int. J. Med. Inf.*, **67**, 49–61.

Friedman,C. *et al*. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17** (Suppl 1), S74–S82.

Fukuda,K. *et al*. (1998) Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput.*, 707–718.

Gavin,A.C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Giot,L. *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Hermjakob,H. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32** (Database issue), D452–D455.

Hirschman,L. *et al*. (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, **18**, 1553–1561.

Hirschman,L. *et al*. (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, **6**(Suppl 1), S11.

Ho,Y. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Hoffmann,R. and Valencia,A. (2003a) Life cycles of successful genes. *Trends Genet.*, **19**, 79–81.

Hoffmann,R. and Valencia,A. (2003b) Protein interaction: same network, different hubs. *Trends Genet.*, **19**, 681–683.

Ito,T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jenssen,T.K. *et al*. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.

Jeong,H. *et al*. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Kim,W. *et al*. (2001) Automatic MeSH term assignment and quality assessment. *Proc. AMIA Symp.*, 319–323.

Koch,C. and Laurent,G. (1999) Complexity and the nervous system. *Science*, **284**, 96–98.

Krauthammer,M. *et al*. (2000) Using BLAST for identifying gene and protein names in journal articles. *Gene*, **259**, 245–252.

Masys,D.R. *et al*. (2001) Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, **17**, 319–326.

Mika,S. and Rost,B. (2004) Protein names precisely peeled off free text. *Bioinformatics*, **20** (Suppl 1), I241–I247.

Morgan,A., Hirschman,L., Yeh,A. and Colosimo,M. (2003) Gene Name Extraction Using FlyBase Resources. *ACL-03 Workshop on Natural Language Processing in Biomedicine*, 1–8.

Motter,A.E. *et al*. (2002) Topology of the conceptual network of language. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, **65**, 065102.

Ono,T. *et al*. (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, **17**, 155–161.

Perez-Iratxeta,C. *et al*. (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.

Pieprzyk,J. and Sadeghiyan,B. (1993) *Design of hashing algorithms*, Springer-Verlag, Berlin, New York.

Proux,D. *et al*. (1998) Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. *Genome Inform. Ser. Workshop Genome. Inform.*, **9**, 72–80.

Pruitt,K.D. *et al*. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.*, **16**, 44–47.

Stapley,B.J. *et al*. (2002) Predicting the sub-cellular location of proteins from text using support vector machines. *Pac. Symp. Biocomput.*, 374–385.

Stuart,J.M. *et al*. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Tanabe,L. *et al*. (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, **27**, 1210–1214, 1216–1217.

Tsuruoka,Y. and Tsujii,J. (2003) Boosting precision and recall of dictionary-based protein name recognition. *ACL-03 Workshop on Natural Language Processing in Biomedicine*, Sapparo, Japan 41–48.

Uetz,P. *et al*. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Van Rijsbergen,C.J. (1979) *Information Retrieval*. 2nd edition ed. Butterworths, London.

von Mering,C. *et al*. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.