Mining the internet for scientific discoveries: what can automatic page tagging tell us about the study of genes?

Shao Chih Kuo[1,2], Andrea Splendiani[1], Michael Defoin-Platel[1], and Chris Rawlings[1].

[1] Department of Biomathematics and Bioinformatics, Rothamsted Research,
Harpenden AL5 2JQ, United Kingdom
[2] School of Computing Science, Claremont Tower, Newcastle University,
Newcastle Upon Tyne NE1 7RU, United Kingdom

shaochih.kuo@bbsrc.ac.uk

Abstract. The internet is not only a platform for publishing documents; it is a provider of data and services. Increasingly, scientific disciplines are exposing their tools and data to the internet, as a result, some scientific problems have become essentially internet mining problems. We show that candidate gene prioritisation, a challenging problem in biology, is essentially an internet mining problem. Thus, improving our ability to mine Future Internet Knowledge Bases (FIKBs) will advance biology and other sciences.

Keywords: Bioinformatics, semantic graph, graph mining, gene prioritisation, automatic page tagging

1 Introduction

The internet has been, and still is, primarily concerned with publishing documents. However, it is clearly also a provider of data and services: scientific data is increasingly accessible on the internet, and many scientific tools are made available via the web, as web services, web applications, or otherwise exposed to the internet.

This is particularly evident in the Life Science domain, which has embraced theinternet as a medium for publishing data and tools. To cite a few examples, for molecular data, the journal *Nucleic Acids Research* has tracked 1,230 databases [1], covering a diverse range of topics and this figure is growing at an rapid rate. Likewise, the *BioCatalogue* directory tracks 1,695 publicly available web services of bioinformatics analysis tools [2]. PubMed, the web's largest bibliography, is also life science centred with a historical focus on biomedical topics. Therefore the internet is, amongst other things, a distributed knowledge base for biological studies where the network of biological entities and their relations is described "in the web": via interlinked websites, or more explicitly, as RDF graphs [3].

In light of the "internetisation" of biological data and resources, we assert that many biological problems are de facto internet mining problems, analogous to more conventional internet mining problems. Therefore improving our ability to mine

Future Internet Knowledge Bases (FIKBs) will certainly advance biology and other sciences. We demonstrate this by showing how the problem of gene prioritisation is analogous to automatic page tagging.

2 Gene prioritisation: a biological problem

Finding causes that influence particular traits is an important challenge in biology; whether it is locating disease genes affecting humans, factors decreasing food production for cereals, or factors increasing industrial insulin production, fundamentally, the goal is the same, to find causes of biological traits. Often the causes under study are genetic actors, and the methods employed to examine them invariably rely on drawing parallels against the body of studied genes; that is to say, given some new gene of study, the assumption is always that it works in a similar way to closely evolutionarily related genes [4].

This assertion underpins the choice to focus study on model organisms, usually organisms which lend themselves to study (i.e. by virtue of having easily observed characteristics or by being cheap to work with) which are representative of their respective classes [5], for example, mouse is commonly used as a model organism for human. For studying a newly discovered gene, bioinformatics can be used first to identify studied evolutionarily related genes by various similarity measures and then to transpose information to the unstudied gene by assigning it putative functions [6].

Using observations and statistical techniques, associations can be drawn between complex traits and genomic regions, however, these regions can be large [7] and also the costs of gene testing may be high, so as to make the cost of exhaustively testing every gene in the region prohibitive. For the biofuels crop willow, biomass is an important trait involved in the production of biofuels. Testing time for a single gene for its influence on this trait ranges from months to years, and genomic regions derived from the statistical techniques may contain several hundred genes. As randomly testing genes is unlikely to reveal trait-affecting genes, this is a clear case for gene prioritisation techniques.

When analysing genes, whilst some useful knowledge may be gleaned from analysing their sequences directly [8], by and large, the bulk of useful knowledge about these genes will be derived from comparing or otherwise associating the newly sequence genes to the corpus of well-studied genes [9], to existing pathways [10], to publications [11], and any other available data. These associations induce a semantically heterogeneous graph, with each gene comparison or association method asserting a new type of relationship between genes from the newly sequenced organism to the wider general body of knowledge, which itself would be a semantic heterogeneous graph (see Figure 1 for an example). Once viewed as a graph, descriptions of complex traits of interest can then be represented as a collection of nodes in the graph representing functional annotations, such as those from the various biological ontologies [12] or controlled vocabularies. Finding good ways to prioritise genes for experimental testing for complex biological traits then becomes equivalent to ranking the overall association in a heterogeneous graph, from a set of nodes of a gene type (genes from the newly sequenced organism) to a set of a set of nodes of the annotation types.

3 Automatic page tagging: an analogy

The pattern of the problem presented earlier is not unique to the biological domain. As an example, the same pattern can also be found in an automatic page tagging system that works by comparing untagged pages against an existing corpus, where pages may be associated by relationships such as: “belonging to the same domain”, “being written in the same language”, or “belonging to the same web ring”. Furthermore, pages can be related by the similarity of their structure, by shared keywords, by a shared audience of the pages, and by other page comparison methods. These relationships may be more or less informative, but will induce a semantically heterogeneous graph.

Suppose that the pages in the existing corpus have been assigned appropriate tags by curation, these tags may be from a controlled vocabulary, ontology, or free (which could still fall into a structure such as WordNet [13]). An automatic tagging method then might be, for each untagged page, to determine the strength of association between that page and the existing tags, and then assign tags according to the strength of association (with some sensible threshold).

Then tag based search (by single tags or by collection of tags) of the new, automatically tagged pages would return an ordered list of those pages most associated to those tags, this order can utilise the semantic distances between tags, and makes the problem analogous to the gene prioritisation problem. An example of a graph that this view of the tagging problem may induce is shown in Figure 1.

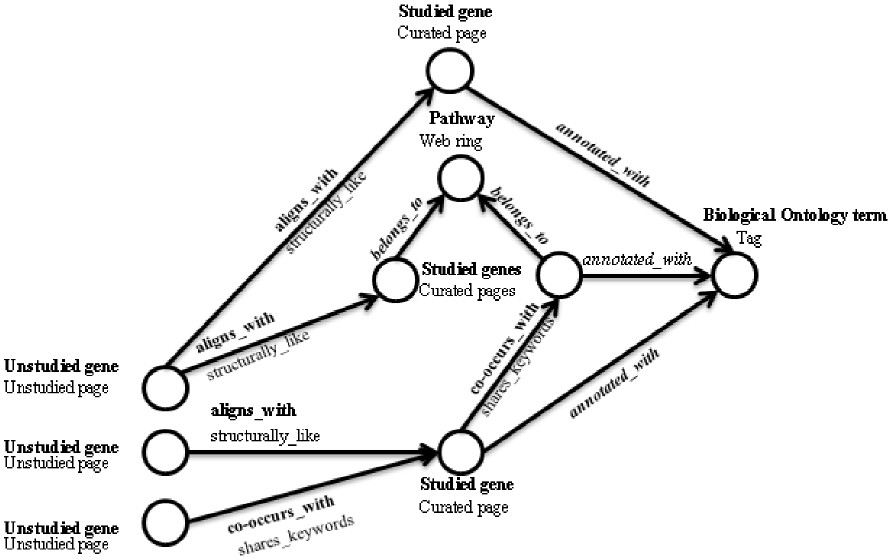

Fig. 1. Two example graphs, candidate gene prioritization based on studied genes, and automatic page tagging based on curated pages sharing the same graph topology. Bold type represent to the “gene version” of this graph topology, whereas regular type represents the “page version” of this graph topology. Where edge/node types are the same in both cases, they are italicised.

4 Gene prioritisation: an internet mining problem

In the previous section, we have shown how a typical bioinformatics problem, gene prioritisation, is analogous to a typical internet mining problem. Beyond this analogy, this and other bioinformatics studies should also be considered internet mining problems in the own right.

Each of the node types shown in the gene prioritisation example in Figure 1 is represented by one or more internet resources, as are the edge types. For each of the node types shown in Figure 1, one source of this type of data is given in Table 1. For each of the edge types shown in Figure 1, one source of data for this type of relationship, or one tool for asserting this type of relationship, is shown in Table 2. Biological entities and relationships are encoded in a variety of forms, as documents, in structured data, and combinations of both, for our purposes, we only wish to illustrate that at least one form is (and in general, many forms are) available on the internet.

Thus, heterogeneous graphs that can be used for solving the candidate gene prioritisation problem are directly available on the internet, and along with other scientific resources, will be part of Future Internet Knowledge Bases (FIKBs).

Nodes	*Source*	*URL*
Genes	Ensembl [14]	http://www.ensembl.org/index.html
Pathways	KEGG Pathway [15]	http://www.genome.jp/kegg/pathway.html
Annotation terms	Gene Ontology [12]	http://www.geneontology.org/

Table 1. Biological entity types and a source of information about them available on the internet.

Edges	*Source*	*URL*
Alignment (aligns_with)	NCBI BLAST [16]	http://blast.ncbi.nlm.nih.gov/Blast.cgi
Annotation (annotated_with)	GOA [17]	http://www.geneontology.org/GO.downloads.annotations.shtml
Pathway associations (belongs_to)	PathExpress [18]	http://bioinfoserver.rsbs.anu.edu.au/utils/PathExpress/
Text mining (co-occurs_with)	PPI Finder [19]	http://liweilab.genetics.ac.cn/tm/

Table 2. Types of relationships between biological data, and examples of internet databases containing such data, or web tools available to generate such data

5 Concluding thoughts

In conclusion, with the greater availability of scientific resources on the internet, tasks in mining scientific data will increasingly become internet mining problems. Scientific research will increasingly rely on the design and availability of dedicated Future Internet Knowledge Bases (FIKBs), and on the development of associated methods to analyse them.

This brings with it, both new challenges and new opportunities. Whilst we have illustrated our case with a problem in the biological domain, the principles hold more widely for other sciences. Some scientific problems have parallels amongst existing internet mining problems, and it is reasonable to expect that advances in techniques in mining the future internet will provide solutions to scientific problems, and vice versa.

Acknowledgements

The authors would like to thank Julia Halder for valuable comments, the authors also gratefully acknowledge funding from the UK Biotechnology and Biological Sciences Research Council (DTG BB/F529038/1, SABR Grant BB/F006039/1).

References

1. Cochrane, G.R., Galperin, M.Y.: The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. Nucleic Acids Res 38, D1-4 (2010)
2. Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., Lopez, R., Goble, C.A.: BioCatalogue: a universal catalogue of web services for the life sciences. Nucleic Acids Res 38 Suppl, W689-694 (2010)
3. Antezana, E., Blonde, W., Egana, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V., Kuiper, M.: BioGateway: a semantic systems biology tool for the life sciences. Bmc Bioinformatics 10 Suppl 10, S11 (2009)
4. Eisenberg, D., Marcotte, E.M., Xenarios, I., Yeates, T.O.: Protein function in the post-genomic era. Nature 405, 823-826 (2000)
5. Hedges, S.B.: The origin and evolution of model organisms. Nat Rev Genet 3, 838-849 (2002)
6. Kaminski, N.: Bioinformatics. A user's perspective. Am J Respir Cell Mol Biol 23, 705-711 (2000)
7. Kleeberger, S.R., Schwartz, D.A.: From quantitative trait locus to gene: a work in progress. Am J Respir Crit Care Med 171, 804-805 (2005)
8. Skolnick, J., Fetrow, J.S.: From genes to protein structure and function: novel applications of computational approaches in the genomic era. Trends Biotechnol 18, 34-39 (2000)

9. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J Mol Biol 215, 403-410 (1990)

10. Dale, J.M., Popescu, L., Karp, P.D.: Machine learning methods for metabolic pathway prediction. Bmc Bioinformatics 11, 15 (2010)

11. Krallinger, M., Valencia, A., Hirschman, L.: Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol 9 Suppl 2, S8 (2008)

12. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25-29 (2000)

13. Sigman, M., Cecchi, G.A.: Global organization of the Wordnet lexicon. Proc Natl Acad Sci U S A 99, 1742-1747 (2002)

14. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., Clamp, M.: The Ensembl genome database project. Nucleic Acids Res 30, 38-41 (2002)

15. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M.: From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34, D354-357 (2006)

16. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., Madden, T.L.: NCBI BLAST: a better web interface. Nucleic Acids Res 36, W5-9 (2008)

17. Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., Apweiler, R.: The GOA database in 2009--an integrated Gene Ontology Annotation resource. Nucleic Acids Res 37, D396-403 (2009)

18. Goffard, N., Weiller, G.: PathExpress: a web-based tool to identify relevant pathways in gene expression data. Nucleic Acids Res 35, W176-181 (2007)

19. He, M., Wang, Y., Li, W.: PPI finder: a mining tool for human protein-protein interactions. Plos One 4, e4554 (2009)