Gene Indexing: Characterization and Analysis of NLM's GeneRIFs

Joyce A. Mitchell, PhD[1], Alan R. Aronson, PhD[2], James G. Mork, MS[2],
Lillian C. Folk, MS[1], Susanne M. Humphrey, MLS[2], Janice M. Ward, MS[2]
[1]University of Missouri – Columbia, [2]National Library of Medicine

Abstract

We present an initial analysis of the National Library of Medicine's (NLM) Gene Indexing initiative. Gene Indexing occurs at the time of indexing for all 4600 journals and over 500,000 articles added to PubMed/MEDLINE each year. Gene Indexing links articles about the basic biology of a gene or protein within eight model organisms to a specific record in the NLM's LocusLink database of gene products. The result is an entry called a Gene Reference Into Function (GeneRIF) within the LocusLink database. We analyzed the numbers of GeneRIFs produced in the first year of GeneRIF production. 27,645 GeneRIFs were produced, pertaining to 9126 loci over eight model organisms. 60% of these were associated with human genes and 27% with mouse genes. About 80% discuss genes with an established MeSH Heading or other MeSH term. We developed a prototype functional alerting system for researchers based on the GeneRIFs, and a strategy to find all of the literature related to genes. We conclude that the Gene Indexing initiative adds considerable value to the life sciences research community.

Introduction

As the quantity of biological information arising from the Human Genome Project grows exponentially, the need to annotate the data with specific literature references becomes more critical. Such annotation has long been the standard of practice within the major protein resources like PIR, PDG, SWISS-PROT and NLM's LocusLink [1-4] . Their strategy has been to update the literature associated with specific sequences, proteins or other gene products with a focus on a gene or sets of genes. However, this strategy is being outstripped by the growth of literature related to protein structure, function and genetics, making it difficult to handle updating the existing information simultaneously with adding new information. It was due to this gap between the growth of knowledge and the difficulty with the literature annotation function that the National Library of Medicine decided to institute "Gene Indexing". After some initial exploration of techniques and pilot projects [5,6], the NLM started April 1, 2002 to prospectively use gene indexing for all articles in all of the 4600 biomedical journals that are included in PubMed/MEDLINE [7].

The NLM Gene Indexing initiative links any article about the basic biology of a gene or protein within eight organisms to a specific record in the NLM's LocusLink database of gene products. The result is an entry called a Gene Reference Into Function (GeneRIF) within the LocusLink database, giving the PubMedID of the article and a short phrase selected primarily from the title or abstract of the article to indicate why this article was chosen to be linked to this LocusLink record.

The purpose of this paper is to present an initial analysis of the LocusLink GeneRIFs. We focus on the initial year of the production of GeneRIFs. We present a characterization of the numbers of GeneRIFs produced and over which species. We also present an initial application of GeneRIFs for the purpose of alerting researchers about literature on gene products. We assess what GeneRIFs contribute to the task of accessing the literature.

Background

NLM's Gene Indexing initiative utilizes the already established MeSH indexing process. MeSH indexers, who index more than 500,000 journal articles per year for PubMed/MEDLINE, create the annotations in the GeneRIF section of LocusLink during the routine process of indexing.

LocusLink [2,8], developed by the National Center for Biotechnology Information (NCBI), provides a single query interface to sequence and descriptive information about genetic loci in eight model organisms: human, mouse, rat, fruit fly, zebrafish, HIV-1, cow and the roundworm. Each LocusLink record contains a variety of information about a gene and its protein products including official nomenclature, aliases, sequence accession numbers, phenotypes, EC numbers, OMIM numbers, UniGene clusters, homologies, map locations and related web sites. The record also contains a section named "Function" which include "References into Function" or GeneRIFs. GeneRIFs provide annotated links between journal articles containing functional information about the gene or protein and the LocusLink record. This makes journal articles containing specific information about a gene and its protein products easily accessible to LocusLink users. By scanning the GeneRIFs, they can identify articles of interest and go directly to the PubMed/MEDLINE record. Also, by using the LinkOut feature, users retrieving articles in PubMed/MEDLINE can go directly to LocusLink for additional information about the genes or proteins discussed in a particular citation.

Journal articles that focus on the basic biology of a gene or its protein products are candidates for linking. Basic biology includes isolation, structure, genetics and function of genes/proteins in normal and disease states. As indexers perform subject analysis of journal articles for Medical Subject Headings (MESH) indexing, they identify genes/proteins appropriate for linking to LocusLink. They then search LocusLink for the gene or protein and, if found, they create a link between that record and the journal article. Indexers als o provide a concise statement summarizing the new information about the gene or protein that is reported in the article. The PubMed ID of the article and the statement appear in the GeneRIF section of the LocusLink record. The indexers also create a supplementary concept record (SCR) if there is no MeSH Heading for the gene product in the article. SCR's are a set of MeSH concepts, mostly chemicals, updated weekly. This whole process occurs within a module of NLM’s web based Data Creation and Maintenance System (DCMS), the system used to create and index journal citations for PubMed/MEDLINE.

The Index Section of NLM began the Gene Indexing initiative as a pilot project in the summer of 2001. Index section staff members were trained and began gene indexing in December, 2001. In March, 2002, contract indexers were trained and began gene indexing in April. Since the beginning of the Gene Indexing project to February 13, 2003, there were 27,645 GeneRIFs created. Indexers now average 600 links per week. Approximately 59% of the links are for human, 27% for mouse, 10% for rat and 3% for the fruit fly, and less than 1% for zebrafish, cow, HIV-1 and roundworm combined (see Table 1).

Although indexers create most of the GeneRifs, the public can also submit GeneRifs. The Internet address is http://www.ncbi.nlm.nih.gov/LocusLink/.

Methods

In order to begin an analysis of GeneRIFs, we downloaded LocusLink files from the NCBI web site as of February 13, 2003. For those LocusLink records having at least one GeneRIF, we extracted several kinds of information: name, summary and detail. Several kinds of names are recorded in LocusLink. Genes can have one or more of the following names: OFFICIAL GENE NAME, PREFERRED GENE NAME, OFFICIAL SYMBOL, PREFERRED SYMBOL and ALIAS SYMBOL. And the products of a gene have names: ALIAS PROT, PREFERRED PRODUCT and PRODUCT. We extracted all names for a record, assigning a preferred name for genes and gene products in the process. Extracted summary information consists of the LOCUSID, the number of GeneRIFs for the locus, and the species of the gene. Finally, detailed information for each GeneRIF consists of LOCUSID, an index (1-N) for the GeneRIF, the number (N) of GeneRIFs for this locus, the PubMed ID, and the summarizing statement, or snippet, justifying the GeneRIF. We computed summary statistics for the extracted data (see Table 1) and for subsets of the data for the most common species: human, mouse, rat and fruit fly.

We were interested in determining the novelty of the genes having GeneRIFs, and we decided that one way to assess novelty would be to compute which genes were represented in NLM's Medical Subject Headings (MeSH). We reasoned that established genes would have MeSH Headings (MHs) and newly reported ones would not have a MH but perhaps have a SCR or entry vocabulary. We used the MetaMap program [9] to map all of the names to concepts in the Unified Medical Language System (UMLS) Metathesaurus, and we then determined which of these concepts were actually MHs and SCRs by using a “Restrict-To-MeSH” module of the MetaMap [10]. We analyzed the remainder to find additional SCRs.

We analyzed the GeneRIFs with respect to whether they were related to genes that cause human diseases. To accomplish this, we ran a query on LocusLink to extract the human genes that cause diseases and with known sequences. This query “has_seq AND disease_known AND organism=human” returned 1,459 loci on February 20, 2003. Of these loci, 1,079 had articles published in the GeneRIF file of our study. Of these, 955 loci had a MH or SCR while 124 had no MeSH terms. We considered retrieval strategies focused on these genes to determine how best to find all relevant information about these genes including from years prior to the initiation of the GeneRIFs.

Because of our interest in keeping current with the literature written about specific genes, we developed a prototype functional alerting system for researchers interested in specific loci and using the GeneRIF files. The alerts are developed from the "temp1" file publicly available for download from LocusLink. In this application, a series of Perl scripts select the loci with the GeneRIFs, analyze those GeneRIFs which have been added since the last update, and send an email alert to the researchers interested in specific loci. In the production system, a WWW page will be provided by which researchers can self-register to receive updates on specific loci, or on all loci for a specific organis m; in addition, researchers will be able to specify the frequency with which they receive alerts for each locus or organism (i.e., monthly, quarterly). We chose to do an alerting system because other applications have shown that this is an effective yet non-invasive method of keeping people informed about new information [[11,12].

Results

Table 1 shows the distribution of the 27,645 GeneRIFs across loci and species in the files studied. Table 1 also gives a breakdown of which of these loci have MHs, those with SCRs and those with neither.

Table 2 shows the number of loci by species with multiple GeneRIFs. Most of the loci have not been the subject of any literature over the past year, and many loci have only a few GeneRIF articles. By contrast, 24 loci have received much attention with over 50 articles written on them.

We investigated a retrieval strategy for genes that would supplement the newly created GeneRIFs and could assist with retrieval of information for the time before the GeneRIFs were available. This was done on an ad-hoc basis. Appropriate MeSH terms for use in searching can be found in NLM's MeSH Browser (www.nlm.nih.gov/mesh/MBrowser.html by entering the gene symbols, full forms of the gene ignoring unimportant words, or the Enzyme Commission (EC) number.

We found optimal retrieval on PubMed/MEDLINE using the union of the gene symbol in the text and the MH/SCR to match MeSH indexing or text phrases. For example: ACTN4 OR "alpha-actinin 4". If the MeSH term is a SCR, then use the regular MeSH term to which the SCR maps during indexing, for example: (ACTN4 OR "alpha-actinin 4") AND Actinin[majr]. To limit retrieval to those citations likely to concern gene biology, this strategy can be further intersected with broad-based biology terms, i.e., (deficiency[sh:noexp] OR enzymology [sh:noexp] OR immunology[sh:noexp] OR metabolism[sh:noexp] OR pathology[sh:noexp] OR physiology[sh:noexp] OR physiopathology [sh:noexp]).

Table 2: Number of Loci and Distribution of GeneRIFs by Species

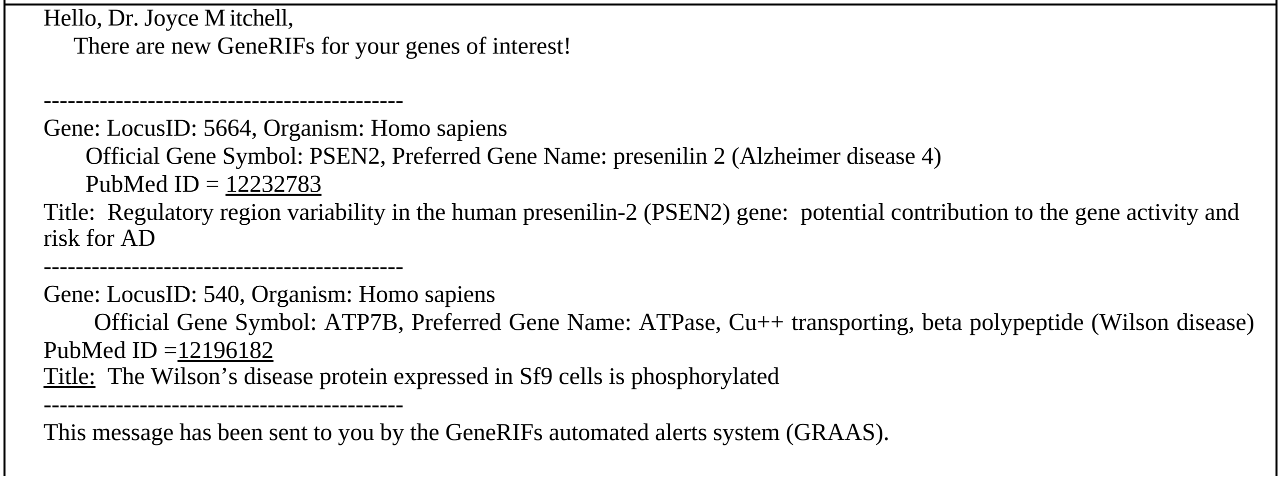
	# of Loci by Species			
# of GeneRIFs	**fruit fly**	**human**	**mouse**	**rat**
0	28,904	31,066	80,854	35,986
1- 5	514	3,875	2,462	1,143
6 - 20	12	540	235	81
21 - 50	1	84	20	2
> 50	0	21	3	0
Total Loci	29,431	35,586	83,574	37,212

Figure 1 shows an example of an email alert related to two new publications related to genes of interest to a researcher. The individual subscribes to the GeneRIF automated alerts system (GRAAS) and indicates by LocusID which genes are of interest to him/her. The GRAAS analyzes the GeneRIFs related to the loci of interest to the researcher and sends an email when new GeneRIFs appear. The example GRAAS email indicates that for two genes (PSEN2 and ATP7B) there are new GeneRIFs discussing those genes. The researcher can then link via the PubMed/MEDLINE ID to the article of the GeneRIF. A future option of GRAAS is to retrieve all information on these genes and will use the retrieval strategy discussed above.

Table 1: Distribution of GeneRIFs across species, MeSH Headings and SCRs

Species	**#GRIFs**	**%**	**#loci**	**#loci w/ MH**	**%**	**#loci w/ SCR**	**%**	**#loci w/o MeSH Terms**	**%**
fruit fly	891	3.2%	527	260	49.3%	200	38.0%	67	12.7%
human	16,267	58.8%	4,521	2,047	45.3%	1,688	37.3%	786	17.4%
mouse	7,401	26.8%	2,721	1,306	48.0%	987	36.3%	428	15.7%
rat	2,866	10.4%	1,227	515	42.0%	383	31.2%	329	26.8%
other	220	0.8%	130	45	34.6%	38	29.2%	47	36.2%
Totals	**27,645**	**100.0%**	**9,126**	**4,173**	**45.7%**	**3,296**	**36.1%**	**1,657**	**18.2%**

Figure 1: Output of Prototype GeneRIF Automated Alerts System (GRAAS)



Hello, Dr. Joyce M itchell,
There are new GeneRIFs for your genes of interest!

Gene: LocusID: 5664, Organism: Homo sapiens
Official Gene Symbol: PSEN2, Preferred Gene Name: presenilin 2 (Alzheimer disease 4)
PubMed ID = 12232783
Title: Regulatory region variability in the human presenilin-2 (PSEN2) gene: potential contribution to the gene activity and risk for AD

Gene: LocusID: 540, Organism: Homo sapiens
Official Gene Symbol: ATP7B, Preferred Gene Name: ATPase, Cu++ transporting, beta polypeptide (Wilson disease)
PubMed ID =12196182
Title: The Wilson's disease protein expressed in Sf9 cells is phosphorylated
----
This message has been sent to you by the GeneRIFs automated alerts system (GRAAS).

Discussion

Almost 60% of the GeneRIFs produced in the first year of the Gene Indexing project are associated with human genes. This is not surprising given the emphasis on the Human Genome Project over the last ten years and the final completion of the genome sequence in April, 2003. It also reflects the content of the 4,600 journals indexed by the NLM that are more focused on humans than on other organisms. In the first year of GeneRIFs, many of these articles focused on studies of human genes that cause disease. The GeneRIFs covered 1,079 disease gene loci, many of which were new genes without a MH but only a SCR. Also evident from the large number of GeneRIFs was the focus on the mouse genome with 27% of the GeneRIFs.

The focus of research activity has obvious "hot spots" with large numbers of publications on some loci. The hottest spots are the TP53 (tumor protein p53) gene in humans with 234 articles written and the homologous protein in the mouse, the Trp53 (transformation related protein 53) gene, with 79 articles. The largest number of articles in the rat was for the Tnf (tumor necrosis factor superfamily, member 2) gene and the Vegf (vascular endothelial growth factor) gene with 22 articles each. In the fruit fly the wg (wingless) gene was the largest with 19 articles.

For those articles with GeneRIFs, 45% discuss gene products that have an established MeSH heading and thus are relatively easy to retrieve. These most likely represent well-known genes. Another 36% have SCR indexing and are usually newer concepts; the SCR terms ease retrieval for experienced searchers or for those using software that automatically maps to the SCR when it exists. The remaining 18% discuss genes without a MH or a SCR; retrieval of literature about these genes provides a challenge for most searchers. A major effect of the GeneRIF project will be to make the retrieval easier for concepts without MeSH headings since the researcher can automatically traverse from the LocusLink record to the relevant literature. This helps to ameliorate the well-known difficulties in retrieving literature about genes due to the unusual and multiple names associated with these newly emerging entities (13).

The GeneRIF link between PubMed/MEDLINE and LocusLink opens up a new array of potential applications that can assist in keeping researchers up to date. We have created a prototype but functional GeneRIF Automated Alerting System (GRAAS) to illustrate this point. This system is similar to SDI (selected dissemination of information) software that alerts researchers about topics of interest when the literature comes into print. The GRAAS project is an SDI application but focused on biological functions that are often hard to retrieve with standard MeSH queries. Other applications will undoubtedly be developed utilizing the GeneRIFs.

This analysis has provided a view of the large amount of literature being produced by the world research community concentrated on the biology of genes. One measure of the literature activity over the past year is to compare the number of GeneRIFS and the number of SWISS-PROT literature references for human genes. Table 3 gives the numbers for this comparison. The SWISS-PROT numbers[1] are the results of several years of curation of protein sequences and associated literature. The GeneRIF numbers are from approximately one year.

[1] SWISS-PROT data from Feb 22, 2003 at http://us.expasy.org/sprot/hpi/hpi_stat.html

Table 3: SWISS-PROT and GeneRIF Comparison

System	Annotated human loci	Literature references (Unique)	Max per entry	Avg per entry
SWISS-PROT	9124	29,129 (23,560)	143	3.19
GeneRIF	4521	16,267 (12,004)	234	3.6

At this rate of GeneRIF production, the number of GeneRIFs will overtake the number of SWISS-PROT references in very short order. This implies that the strategy of Gene Indexing is a method by which the linkage of new literature references to specific gene loci is more complete and more timely than the selected approaches used previously. And because LocusLink is interlinked with multiple bioinformatics databases world wide, the GeneRIF project effectively adds literature citation links to many more protein databases than just LocusLink, including the SWISS-PROT, PDB, PIR systems and others.

Conclusion

We analyzed the first year of production of the NLM's Gene Indexing initiative. Gene Indexing results in a GeneRIF in the LocusLink record related to a specific literature reference. We conclude that Gene Indexing adds considerable value to the life sciences research community:

- Specific links between the PubMed/ MEDLINE files and gene products within LocusLink make it easier for researchers to find literature written about specific loci. With half of the loci having no MeSH Heading, precise retrieval of relevant literature is often difficult. The GeneRIF project eases this burden.
- Global coverage for all journal literature indexed by the NLM and all loci in the eight model species is an effective strategy for maintaining currency with the literature.
- GeneRIFs provide a new avenue for applications development. One avenue we pursued was to create an application using the GeneRIFs to assist researchers to keep current on loci of interest. Undoubtedly other applications will be developed to assist the research community.
- It may be possible to further develop the automated systems to interactively assist indexers in creating GeneRIFs or to retrieve relevant citations for adding GeneRIFs retroactively, thus making the production of GeneRIFs more efficient.

Acknowledgements

This research was supported in part by NLM's Biomedical and Health Informatics Research Training grant 2-T15-LM07089-11.

References

(1) Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research 2003; 31(1):365-370.

(2) Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU et al. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Research 2002; 30(1):13-16.

(3) Sussman JL, Abola EE, Lin D, Jiang J, Manning NO, Prilusky J. The protein data bank. Bridging the gap between the sequence and 3D structure world. [Review] [15 refs]. Genetica 1999; 106(1-2):149-158.

(4) Wu CH, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu ZZ et al. The Protein Information Resource: an integrated public resource of functional annotation of proteins. Nucleic Acids Research 2002; 30(1):35-37.

(5) Sorden NN, Chang HF, Nelson SJ. Automated indexing of gene symbols. 2001: http://www.nlm.nih.gov/mesh/gene.html.

(6) Ward J. Gene Indexing. NLM Tech Bulletin 2002; 328:e6.

(7) MEDLINE Fact Sheet, Sept 18, 2002. http://www.nlm.nih.gov/pubs/factsheets/medline.html

(8) Pruitt KD, Tatusova TA, Maglott DR. NCBI Reference Sequence project: update and current status. Nucleic Acids Research 2003; 31(1):34-37.

(9) Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap Program. Proc AMIA Symp 2001:17-21.

(10) Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp 1998:815-9.

(11) Balas EA, Austin SM, Mitchell JA, Ewigman BG, Bopp KD, Brown GD. The clinical value of computerized information services. A review of 98 randomized clinical trials. Archives of Family Medicine 1996; 5(5):271-278.

(12) Szilagyi P, Vann J, Bordley C, Chelminski A, Kraus R, Margolis P et al. Interventions aimed at improving immunization rates. Cochrane Database of Systematic Reviews 2002;(Issue:4):4.

(13) Hanisch D, Fluck J, Mevissen HT, Zimmer R. Playing biology's name game: identifying protein names in scientific text. Proceedings of the Pacific Symposium on Biocomputing 2003;403-414.