

Methods for Large-Scale Mining of Networks of Human Genes

Tor-Kristian Jenssen^{1, †}, *Lisa M.J. Öberg*^{2, †},
*Magnus L. Andersson*², and *Jan Komorowski*¹

Abstract

In molecular biology there is much interest in various types of relationships between genes. Due to the complexity and rapid development of this field, much of this knowledge exists only in free-text form. A database of relationships between genes may allow background knowledge to be used in computerised analyses. As far as we know, no comprehensive manually curated database of this kind exists, and constructing and maintaining such a database manually would be very labour-intensive. Efficient automated methods for extraction and structuring of relationships between genes from free-text would be valuable. A database named PubGene has previously been created and it contains a comprehensive network of human genes created by automated extraction of co-occurrence of gene terms in over 10 million MEDLINE records. Co-occurring genes were linked together under the hypothesis that two genes will co-occur only if they have some biological relationship. In this paper, we show that for the subset of human genes encoding enzymes, pairs of co-occurring enzyme genes are significantly more closely related biologically than when these genes are compared randomly. Manual inspection, however, shows that some of the links in PubGene are not correct and it also indicates how the noise can be reduced. We propose a complementary method for automated extraction of relationships between genes by use of information from the Science Citation Index (SCI) database. We relate two genes if they have been co-referred, that is, having reference articles being co-cited in a third article. The alternative approach confirms relationships found in PubGene, and it also finds other relevant relationships. Although further experiments are

¹ Knowledge Systems Group, Department of Computer and Information Science, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.
tor-kristian.jenssen@idi.ntnu.no, jan.komorowski@idi.ntnu.no.

² Molecular Biology, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden.
lisa.oberg@astrazeneca.com, magnus.l.andersson@astrazeneca.com.

[†] These authors contributed equally

required for the SCI approach, the results are encouraging. Furthermore, the two methods combined can be used to generate networks that have high specificity or high sensitivity by either requiring that relationships should be found by both methods or by only one, respectively.

Keywords: Network of genes, text mining, information extraction, term co-occurrence, article co-citation.

1. Introduction

New discoveries in biomedicine are published at an enormous and constantly increasing pace. The MEDLINE citation database contains citations including titles and abstracts of papers from more than 4,000 scientific journals. MEDLINE contains over 10 million citation records for articles from 1966 to present. The number of indexed articles from the year 1991 is 386,866, while for 1999 the number has increased to as many as 446,178. The growth in number of publications related to molecular biology and genetics is even more impressive. A search in MEDLINE for publications matching 'gene' gives 27,372 articles from 1991, while for the year 1999, the same search returns 52,603 articles.

Knowledge of gene interactions and how genes are related to each other are very important. Such relations may be, for instance, physical interactions between the encoded proteins, regulatory interactions, chromosomal co-localisation, homology, or other kinds of similarity. Given the large number of new publications together with the complexity of this field, it is a challenge to keep updated on new results, even for a handful of genes. Estimates predict the total number of human genes to be somewhere between 30 and 100 thousand. It is obvious that efficient information retrieval and information extraction will become very important in order to cope with the large amount of information, and to enable scientists to efficiently make use of existing knowledge. Automated methods that extract and structure information will first of all make knowledge more easily accessible. Secondly, a structured knowledge representation enables computerised methods to make use of this information. Information extraction may thus give a vital contribution in the interpretation of data from high-throughput gene-expression analyses, which with current technologies allows experiments including tens of thousands of genes.

The PubGene network of human genes [4, 5] is one example of an approach to large-scale extraction of information. The PubGene network was constructed from co-occurrences of genes in the title or abstract of MEDLINE citation records. The underlying hypothesis is that two genes appear in the same abstract only if there is some biological relationship between the two genes. To include most of the identified human genes, a database containing gene nomenclature information was compiled by gathering data from the HUGO Nomenclature Committee³, LocusLink⁴, The Genome Database⁵, and GENATLAS⁶. The nomenclature information includes primary symbols [16], gene names, and literature aliases. The key identifier of a gene is its unique primary symbol. Official gene symbols follow the guidelines set forth by the HUGO Nomenclature Committee. According to these guidelines, a gene symbol should consist of a capital letter followed by letters (preferably capital) and possibly Arabic numerals. Gene symbols are

³ <http://www.gene.ucl.ac.uk/nomenclature/>

⁴ <http://www.ncbi.nlm.nih.gov/LocusLink/>

⁵ <http://www.gdb.org/>

⁶ <http://www.citi2.fr/GENATLAS/>

generally short, typically from 2 to 6 characters, although some longer symbols exist. Gene names can be rather short and specific, e.g. 'insulin', or rather long as for example 'IQ motif containing GTPase activating protein 2'. The network used in this paper contains 14,961 distinct genes. Additional gene symbols will be defined along with the discovery of novel genes. As a consequence, the number of genes in PubGene will increase with future updates. All references to genes will be by the official gene symbol as defined by the HUGO Nomenclature Committee, except for the cases where this committee has not yet included the gene in the nomenclature. Such cases will be explicitly mentioned. Detailed information on a particular gene can be found on the HUGO Nomenclature Committee website or by search from this site.

Gene associations for PubGene were identified by an automated indexing procedure where all MEDLINE records, titles and abstracts, from 1966 to present⁷ were scanned for occurrences of gene symbols and short gene names. Wherever a symbol or a name for a gene was found, this was noted as a match for the gene associated with the symbol or name. The short gene names used were those that consisted of a single word, possibly followed, by a variant designation, such as, for instance, 'insulin', 'cadherin 1' and 'cadherin 2'. In total, more than 10 million MEDLINE records were used in the construction of PubGene. When two genes were found together in an article record, they were linked together and in the end each link was given a weight equal to the number of times the specific pair of genes had co-occurred. As an example, Figure 1 shows the PubGene network surrounding the gene PPARA. Due to limited amount of space, the algorithms used to construct PubGene are not described in more detail in this paper. For more information, see [4, 5] or contact the authors.

The approach used in the construction of PubGene may be characterised as simplistic. As a consequence, the implementation is efficient and has scaled well to make processing of all of MEDLINE citation records feasible on a single PC within a few days of computation time. More sophisticated approaches to term detection, using natural language or statistical models do not scale as well, and, to the best of our knowledge, PubGene is the only database of this kind at this level of comprehensiveness. The PubGene network is already in full scale, in the sense that it includes most identified human genes and is based on almost all articles in MEDLINE. One of the primary uses of the PubGene database is as a summary of the published literature. For this purpose, or as a foundation for analysis of gene-expression data, it is important that the extracted associations are biologically relevant, or at least that the signal to noise ratio is sufficiently high. It is also important that all essential relationships between genes are reflected in the resulting network.

In this paper, we report a comprehensive study of the nature of gene associations found in the PubGene database and in particular, we propose and assess notions of *correctness* and *completeness*. Precision and recall are standard measures for evaluation in information retrieval. In order to avoid confusion with the standard definitions of these terms, we will rather use the related concepts of correctness and completeness in the evaluation of the PubGene network. We will not provide formal definitions of these terms, but rather use them in a general sense. Intuitively, the network can be said to be correct if all the associations are biologically relevant and complete if all biologically relevant associations are present. We present a three-way evaluation of the PubGene network. First, a partial manual evaluation with the use of detailed background knowledge on a small subset of genes is given, see Section 2. Then, in Section 3, we use the enzyme

⁷ As included in MEDLINE by November 1999. Note that only MEDLINE records from 1975 or later contain abstracts.

checked if the abstracts actually referred to the genes in question, and then if this was the case, whether there was a biological relationship. The two genes were peroxisomal proliferator-activated receptor alpha, with the symbol PPARA, and peroxisomal proliferator-activated receptor delta, with the symbol PPARD. These two genes were selected due to prior knowledge on the PPAR gene family. The PPAR genes encode transcription factors involved in the regulation of storage and catabolism of dietary fats [17]. There is also a third PPAR gene, namely PPARG, but this gene was not included in the analysis as we expected it to behave similarly to the other PPAR genes.

The PPARA and PPARD genes have 213 and 41 neighbours, respectively, in PubGene. By reading abstracts we determined whether the association was correct in the sense that both genes were actually referred to in the text. When that was true and the abstract did not give sufficient information to determine the type of relationship, the full text of the articles, when available, was consulted. For PPARD we examined all neighbours, while for PPARA we examined all neighbours with weight higher than 1 as well as a selection of the neighbours with weight 1. The results are shown in Table 1.

<i>Category of PubGene neighbour</i>	<i>PPARA</i>		<i>PPARD</i>	
	<i>1</i>	≥ 2	<i>1</i>	≥ 2
<i>Association weight</i>				
Correct	13	58	8 (7)	6 (5)
Partially correct	8	4		
Incorrect	6	29	21 (9)	6 (3)
Synonym ambiguity, to other gene	3	14	3	3
Synonym ambiguity, to other concept	2	13	18	3
Other	1	2		
Sum	27 (122)*	91	29 (16)	12 (8)

Table 1. *This table gives some statistics on PubGene neighbours of PPARA and PPARD. For both genes, neighbours with weight 1 and neighbours with weight ≥ 2 were examined separately. PPARA has 122 neighbours with weight 1 and as it was considered too time-consuming to examine each one, 27 sample neighbours were randomly selected (*). Each of the PubGene neighbours investigated was categorised as ‘Correct’, ‘Partially correct’ and ‘Incorrect’. The category ‘Incorrect’ was further broken down into sub-categories explaining the reason of incorrectness in more detail. The ‘Partially correct’ category only relates to the PPARA gene. Genes in this category were found mentioned with the PPARG gene, but incorrectly associated with the PPARA gene. Numbers in parenthesis for the PPARD gene relates to the situation where the NUC1 cases are removed from the comparison.*

First of all, we would like to point out that in all cases where the symbols or names actually referred to the correct genes, there also existed a relevant underlying biological relationship. We see that for the PPARD gene, only 14 out of 41 (34%) of the neighbours were correct. Note, however, that as many as 15 of the 27 wrong associations to PPARD could be traced back to the ‘NUC1’ synonym. ‘NUC1’ or ‘NUC-1’ has also often been used to refer to a particular type of cell-line. Moreover, only 2 of the correct associations were found with ‘NUC1’ referring to PPARD. Thus, by disregarding this synonym very much of the noise can be removed, while still keeping almost all of the correct associations.

Many other incorrect associations were caused by gene symbols that are equal to other common abbreviations. For instance, ‘DR-1’, and ‘DR-5’, have been used to abbre-

viate ‘direct repeat 1’ and ‘direct repeat 5’, as well as been used as gene symbols. Incorrect associations are also extracted because several gene symbols have been used for several different genes. As the current algorithm is not capable of resolving the ambiguity, occurrences of such symbols are mapped to all the genes that are associated with the symbol. We see that the extent of this type of incorrect associations is comparable to the other category of symbol ambiguity.

It is interesting to see that the proportion of incorrect (or partially incorrect) associations is considerably lower for associations with higher number of co-occurrences (2 or more) than for the associations that were found only once. For PPAR_D, the improvement is from 0.72 to 0.50, or from 0.53 to 0.38 if we exclude the associations picked up by the ‘NUC1’ symbol. For PPAR_A, the improvement is from 0.52 to 0.36. This means that we can use the weight of the association as an indicator of the expected correctness.

3. Estimation of correctness using enzyme classification

Manual inspection of all gene associations in PubGene, as described in Section 2, was not considered feasible. In order to estimate correctness on a larger scale, we wanted to investigate whether two genes directly linked in PubGene tend to be more closely related biologically than random pairs of genes. Unfortunately, there is no complete database with ‘biological distance’ for all human genes. In fact, had such a database existed, much of the purpose of constructing PubGene would have vanished.

Enzymes are proteins that catalyse chemical reactions. A comprehensive classification does exist for this large class of proteins, namely the Enzyme Classification system⁸. Since, essentially, there is a one-to-one correspondence between the proteins and the genes that code for the proteins, the enzyme classification system can be used to assess the biological relatedness of genes that code for enzymes. The enzyme classification system has existed for a long time, it is well established among scientists, and contains few ambiguities. Enzymes are hierarchically classified into 6 broad categories, which are further subdivided into finer subgroups on 3 levels. EC- (Enzyme Commission) numbers thus consist of 4 numbers, where at each level, the number denotes to which category the enzyme belongs. A review of the development of the enzyme nomenclature and classification can be found in Tipton and Boyce [14]. For future reference, we will make a few definitions. Genes that are directly linked in the PubGene network are called *PubGene neighbours*. A gene with an associated EC-number is called an *enzyme gene* and two enzyme genes that are PubGene neighbours are called *enzyme neighbours*.

In order to be able to see whether enzyme relationship is reflected in the PubGene network, we introduce a notion of distance between enzymes. The hypothesis is that genes encoding enzymes and that are also neighbours in PubGene will have, on average, shorter enzyme distance than random pairs of enzymes. We define the *enzyme distance*, $d(ec_a, ec_b)$, between two EC-numbers ec_a and ec_b as follows. Assume $ec_a = a_1.a_2.a_3.a_4$ and $ec_b = b_1.b_2.b_3.b_4$. Let i_{min} be the smallest i such that a_i is different from b_i , and if $ec_a = ec_b$, let $i_{min} = 5$. Then let $d(ec_a, ec_b) = 5 - i_{min}$. As an illustration, Table 2 shows pairs of enzyme genes, their EC-numbers, and the corresponding enzyme distances.

To extract EC-numbers for genes encoding enzymes, we used the protein databases SWISSPROT and TrEMBL. For each primary gene symbol in PubGene, we searched the Gene Name field of these databases to link to a protein, if found. Then, from the Description field we extracted the EC-number(s), if any. Each gene that encodes a multi-

⁸ <http://www.chem.qmw.ac.uk/iubmb/enzyme/>

functional enzyme was assigned all EC-numbers of the enzyme, thus giving rise to multiple enzyme distances. Next, all enzyme neighbours in PubGene were identified and the associated enzyme distance(s) for each pair was determined. To get the distribution of enzyme distances between enzyme genes in general, all enzyme genes in PubGene were combined in all possible ways. In both cases, a gene pair was considered only once, and genes were not compared to themselves. For genes with multiple EC-numbers, each EC-number was treated separately, and therefore there may be more than one enzyme distance between a pair of enzyme genes.

Enzyme pair	EC 1	EC 2	Enzyme distance
FUT1-FUT2	2.4.1.69	2.4.1.69	0
NDUFB3-CRYZ	1.6.5.3	1.6.5.5	1
TPI1-HSD3B1	5.3.1.1	5.3.3.1	2
ENPEP-BST1	3.4.11.7	3.2.2.5	3
IARS-AHCY	6.1.1.5	3.3.1.1	4

Table 2. Examples of pairs of enzyme genes and the corresponding enzyme distances.

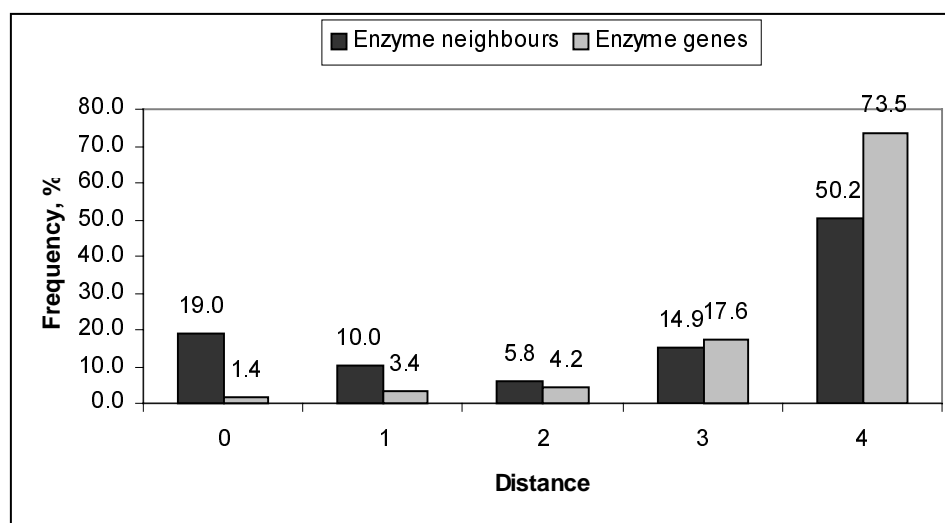


Figure 2. Distributions of enzyme distances between enzyme neighbours and between enzyme genes in general are given in this figure. The two distributions are significantly different, $p < 10^{-215}$. The numbers of enzyme distances in the distributions are 9,735 and 1,197,253, respectively.

In PubGene we identified 1,438 (10%) enzyme genes and in total there were 1,548 EC-numbers associated with those genes. Out of these enzyme genes, 947 (66%) have one or more enzyme neighbours. Figure 2 shows the enzyme distance distributions. We can see that there is an overrepresentation of shorter enzyme distances among enzyme neighbours compared to the random distribution and the two distributions are significantly different ($p < 10^{-215}$ on a χ^2 -test). Comparing the average distance, the average taken over all enzyme pairs was 3.59, while the average over the enzyme neighbours found in PubGene was 2.67. As the background distribution is known, the probability of observing such an average can be computed by assuming that the mean is normally distributed. This

gives a z-statistic of less than -108 , which is extremely significant (the two-tailed p-value of observing $|z| > 5$ is less than 10^{-6}). Also, if we look at the probability of finding a very closely related pair of enzymes, that is a pair where the distance is 0 or 1, the expected probability of such a pair is 0.048. This gives an expected number of 0- or 1-distance pairs among 9,735 of 456, while the actual number is 2,823. Using the normal approximation to the binomial distribution we get a z-statistic of over 113, which is also extremely significant.

Closely related genes that belong to the same protein family are often mentioned together in articles. This is the case, for example, with the genes PPARA, PPARB, and PPARC in the peroxisome proliferator-activated receptor family. As enzymes from the same enzyme family often have the same enzyme classification, this may explain why enzyme neighbours have such an overrepresentation of enzyme distance 0 compared to enzyme pairs in general. To better understand the relationships between enzyme neighbours, we further investigated enzyme genes to see what kind of enzyme neighbours they had. Details for two example genes are shown in Table 3. The example genes are FUT1 and NDUFA5, which we chose to show because they have many enzyme neighbours. FUT1⁹ has 125 neighbours in PubGene out of which 26 are enzymes (21%). As can be seen in Table 3, the enzyme distances from FUT1 to other FUT-enzymes are shorter than the enzyme distances to other enzyme neighbours. A similar pattern can be observed for the other example, NDUFA5¹⁰, which has 26 enzyme neighbours (of 120 in total). Because NDUF-enzymes have two EC-numbers, the number of associated enzyme distances is larger. The fact that close relationship between genes was reflected in the PubGene network supports our hypothesis that the network is biologically relevant.

Enzyme distance	<i>FUT1 to FUTs</i>	<i>FUT1 to others</i>	<i>NDUFA5 to NDUFs</i>	<i>NDUFA5 to others</i>
0	1	-	12	-
1	5	2	-	-
2	-	-	12	-
3	-	11	-	6
4	-	7	-	30

Table 3. Distributions of enzyme distances from *FUT1* and *NDUFA5* to their enzyme neighbours in PubGene. In each case, the distributions for closely related enzymes, 6 *FUT*- and 6 *NDUF*-enzymes, respectively, have been separated from the rest of the enzyme neighbours.

4. Assessment of completeness using SCI citation data

In theory, albeit not very practical, a complete assessment of correctness of PubGene could be made by examination of all the gene associations that are extracted. Since there is no complete database to compare with, completeness is even more difficult to assess. Obviously it is easier to examine whether pairs that are given are correct than to identify associations that are not present. The fact that the kind of gene relationships reflected is rather vague makes the task even harder. Therefore, our approach was to build an alternative network of gene relationships, and then to compare that network to the one in

⁹ Traced to SWISSPROT FUT1_HUMAN, EC 2.4.1.69.

¹⁰ Traced to SWISSPROT NUFM_HUMAN, EC 1.6.5.3 and EC 1.6.99.3.

PubGene. We used reference article data from SWISSPROT and TrEMBL and citation data in the Science Citation Index (SCI) from the Institute for Scientific Information (ISI)¹¹ to define an alternative relation between two genes. This relation was then used to get an estimate of the completeness of PubGene.

SCI is an indexed bibliographic database covering a broad range of scientific journals. For each article, the database stores, for example, the name of the author(s) and the list of cited references. Therefore, it is also possible to get a list of all articles that cite a specific work. We exploited the fact that SCI could be viewed as a network of journal articles, where articles are connected through bibliographic citations. Essentially, what we did was to establish links between genes in PubGene and appropriate articles. First, we defined an *original article* of a gene as an article that was found in the SWISSPROT or TrEMBL entry that was found by search with the primary gene symbols in the Gene Name field in those protein databases. Then, if SCI could be used to find an article that cites at least one of the original articles of two genes, the two genes were defined to be *SCI neighbours* and they were also said to be *co-referred*, see also Figure 3.

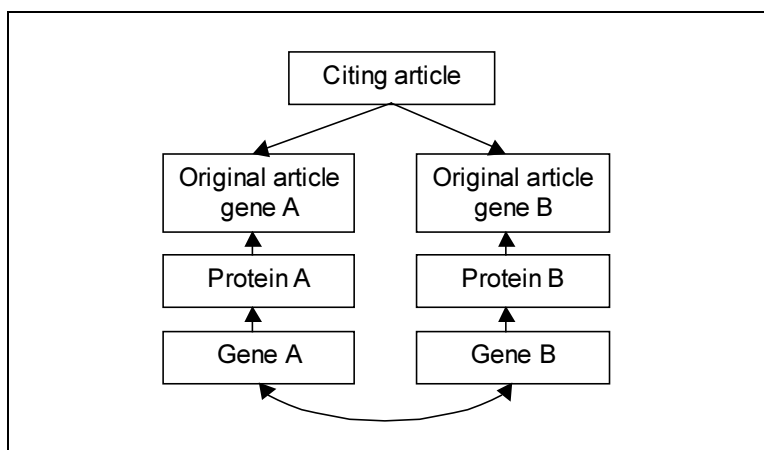


Figure 3. A schematic drawing of how the SCI associations between genes have been established. For a gene, the corresponding protein is identified and an original article of that gene is an article found in the record from the SWISSPROT or the TrEMBL databases. SCI is then used to find articles that cite more than one original article. Whenever such an article is identified, an association is established between the genes behind those original articles.

Through SWISSPROT and TrEMBL we could find original articles, one or more, for 6,579 (44%) of the genes in PubGene. Not all of the genes in PubGene could be traced to a SWISSPROT/TrEMBL record and not all of the protein records contain references to articles. Most proteins have more than one original article, and the total number of such articles was 15,577. We only had access to SCI through a web-browser and therefore, it was not feasible to construct a complete network of SCI neighbours. For a partial estimation of completeness, we chose to create SCI networks for the PPARA and PPARG genes, the same genes that were used in Section 2. As we will show shortly, a large number of the SCI neighbours of PPARA and PPARG cannot be found as PubGene neigh-

¹¹ <http://www.isinet.com>

bours. In order to compare the relation of being linked as SCI neighbours with that of being linked by the PubGene network, we will introduce a second measure of distance, based on PubGene associations. We define the *PubGene distance* as the minimum number of links that has to be crossed in the PubGene network to get from one gene to the other. Thus, the PubGene distance between 'Gene A' and any of its direct neighbours is 1, and from 'Gene A' to a neighbour of a neighbour is 2, and so on. If it is not possible to get from a gene to another by traversing links in the network, we define the PubGene distance as infinite. This notion of distance should not be confused with the previously defined enzyme distance.

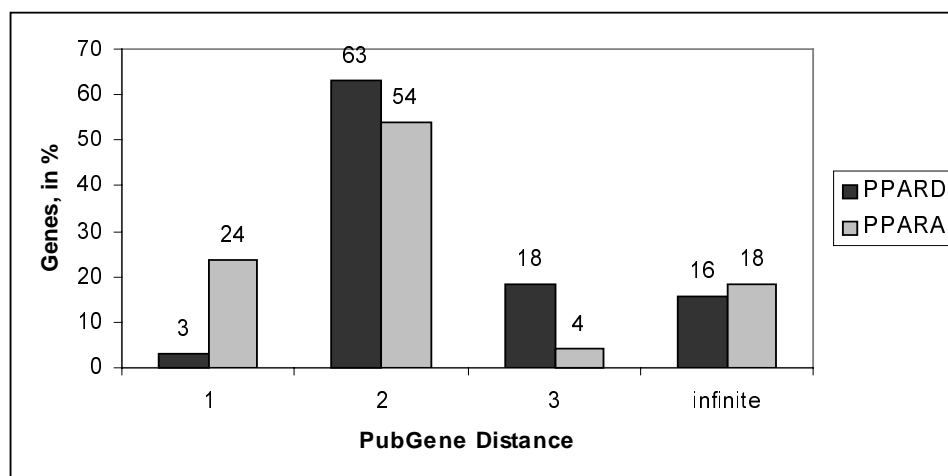


Figure 4. The distributions (in percentage) of PubGene distance from PPARA and PPARD, respectively, to the SCI neighbours of the same. PubGene distance 1 means that the SCI neighbour was also found as a direct neighbour in PubGene. The total numbers of SCI neighbours were 119 for PPARA and 187 for PPARD.

PPARD only has one original article¹². According to SCI there are 198¹³ articles that refer to this article, but we could only obtain reference lists for 154¹⁴ of them. In these 154 reference lists, we identified 187 unique original articles and each of them represents one gene symbol that has been co-referred with the original article of PPARD. Of the 187 genes identified, only 6 (3%) were found at PubGene distance 1 from PPARD. For PPARA, which has two original articles¹⁵, an identical procedure was used to identify 119 SCI neighbours. Figure 4 shows how the 187 SCI neighbours of PPARD and the 119 SCI neighbours of PPARA are distributed over PubGene distance from PPARD and PPARA, respectively. For both genes, PubGene associated fewer genes directly than were related to them through SCI. But, as Figure 4 shows, 84% and 82% of the SCI neighbours of PPARD and PPARA, respectively, are found at PubGene distance 1, 2, or 3.

¹² Schmidt A., *et al*, *Mol. Endocrinol.*, **6**:1634-1641 (1992).

¹³ In May 2000.

¹⁴ The remaining ones are published earlier than 1995.

¹⁵ Sher T., *et al*, *Biochemistry* **32**:5598-5604 (1993); Mukherjee R., *et al*, *J. Steroid Biochem. Mol. Biol.* **51**:157-166 (1994).

We further analysed the SCI neighbours of PPARA and PPARD to see what caused them to be co-referred with PPARA and PPARD, respectively. Table 4 shows how these genes fall into 14 categories, the PubGene distance is also considered. The choice of categories is based on the kind of genes that appear as SCI neighbours of PPARA and PPARD and also on the reason of co-citation. The first 9 categories refer to genes that have a rather specific relationship to PPARs and the following 3 categories refer to genes that have a more general relation to PPARs. Genes with relationships to PPARs that are harder to explain, were referred to the group ‘Other’.

<i>Category of SCI neighbour</i>	<i>PPARA</i>				<i>PPARD</i>			
	<i>1</i>	<i>2</i>	<i>3</i>	<i>Inf.</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>Inf.</i>
Another PPAR	2				2			
Expression is regulated by PPAR	4				3	2		1
Regulates expression of PPAR						2		
Interacts physically/functionally with PPAR	3	3			1	6		
Co-activated/regulated with PPAR						1	2	
Binds a similar response element as PPARs			1			1		
Close chromosome localisation with PPARD						11	5	3
Peroxisomal	2	3				5		
Fat metabolism	1	6		2		17	4	4
NHR, transcription factor	12	20		4		36	5	6
Expression is regulated by other NHR		2				3		
Interacts with NHR	3	12	2	15		12	5	13
Refer to a method, not a gene				1				1
Other	2	16	2	1		22	12	2

Table 4. *The SCI neighbours of PPARA and PPARD have been categorised into 14 groups depending on the kind of relationship these genes have to the PPAR genes. Furthermore, the table indicates how the PubGene distances affect the distribution.*

As Table 4 shows, many of the genes have a quite specific relation to the PPAR gene. For example, PPARA heterodimerises with RXRA [1], and the PPARA/RXRA dimer regulates the expression of the CAT gene. PPAR/RXR dimers further regulate the expression of other genes such as, the UCP2 and the CD36 genes [17]. UCP2 is found at PubGene distance 2 from both PPARA and PPARD and CD36 at PubGene distance 2 from PPARD. The gene WISP2 is co-regulated with PPARD [3] but in PubGene it is found at infinite PubGene distance from PPARD. Among the SCI neighbours of PPARD we also found a large number of genes that are found in the same chromosomal region as PPARD [15], and these genes are found at various PubGene distances, see Table 4. Examples of genes localised close to the PPARD gene are the MAPK11 and TULP1 genes that are found at PubGene distance 2 from PPARD, the ZNF76 and WNT2B genes at PubGene distance 3, while the SFRS3 and RPL10A genes cannot be found by traversing links in the PubGene network.

Among the SCI neighbours we also observe many genes that have a less specific relationship to the PPAR genes. We often observe co-references to other nuclear hormone receptors, NHRs, for example, the ESRRA gene [7], and to genes that have a relationship

to other NHRs as, for example, the co-activator TIF1 that interacts with ESR1 ('estrogen receptor 1') [6] or the gene PS2¹⁶ that is regulated by the ESR1 [6]. The co-reference method can incorrectly associate genes when the original article is cited not because of the gene, but because of a method described in the article. However, this does not occur very frequently, for PPARA and PPARD only one such example was found. The original article of the POR gene is cited in an article about expression of the PPARA gene [8] because it describes a method of an RNase protection assay and not because there is a relation between PPARs and POR.

Among the 187 SCI neighbours of PPARD there are 29 (16%) that have infinite PubGene distance to PPARD. The reasons of co-reference of these 29 genes and PPARD do not differ from the reasons discussed above. For example, there are several genes that code for NHRs and co-factors of NHRs. It is also interesting to note that there are genes that function in the fatty acid β -oxidation, a part of the degradation of fat, because PPARs are key regulators of the fat metabolism. The gene EPHX2 discussed above is also found in this group with infinite PubGene distance. The same observation can be made for PPARA where 22 (18%) of the SCI neighbours have infinite PubGene distance to PPARA.

Further investigation of the PubGene neighbours of PPARA and PPARD shows that at the same time as the SCI network contains many relevant associations that are not found by PubGene, the same observation can be made for PubGene versus SCI. This means that the two methods together are able to find a large number of possibly related pairs of genes that are not found by either method alone. Furthermore, looking at the intersection of neighbours of PPARA and PPARD from PubGene and SCI, we see that there are 28 and 6 genes, respectively, that are found by both methods. We have seen that both methods alone connect pairs of genes whose biological relationships are either very weak or non-existing. Therefore, it is interesting to notice that out of the genes that are found to be related to PPARA and PPARD by both methods, each one have a relevant biological relationship to the PPAR gene.

5. Discussion

The PubGene network of 14,961 human genes is based on co-occurrence of gene terms in more than 10 million MEDLINE records. Our results show that although there is a considerable amount of noise present, the underlying hypothesis that two genes that have been mentioned together also are biologically related is clearly supported. Occasionally, two genes may be mentioned together without a clear biological relationship, but this is rare. However, negative results are sometimes reported, such as, for instance a statement like 'Gene A does not upregulate expression of Gene B'. Nevertheless, in order for such a statement to be put in an abstract, at some level there would usually be some kind of relationship between the two genes.

Stapley and Benoit [11] reported a similar network of *Saccharomyces cerevisiae* (yeast) genes. There are two important differences between their work and the PubGene database. First, the genome of the yeast organism is much better characterised than the human genome, and therefore the nomenclature of yeast genes is more specific and it contains fewer ambiguities. Another difference compared to PubGene is the comprehensiveness since Stapley and Benoit reported a prototype system applied to a small set of 2,524 MEDLINE records published between 1997 and 1998. More sophisti-

¹⁶ This symbol is from the Genome Database.

cated attempts to extract specific types of relationships between genes or proteins have also been reported. These approaches have typically used some kind of natural language processing [10, 13, 9]. So far, these methods have not scaled as well and are likely to be vulnerable to the same problems as PubGene when it comes to extraction of gene names. These methods will also have to resolve the problems related to ambiguous gene symbols and names.

The manual inspection of sample genes in the PubGene network revealed that ambiguous gene symbols caused the PubGene network to include noise in the form of erroneously associated genes. To determine to which of several genes an occurrence of a gene symbol should be assigned is a difficult problem. Incorrect links can also be generated when a gene symbol is identical to an abbreviation used in another context. One likely problem with extraction of gene names is that different authors may write one and the same gene differently. For example, the gene peroxisome proliferator-activated receptor gamma is also referred to as PPARG, PPAR gamma, PPAR-gamma. This causes obvious problems for any algorithm relying on string matching. Another problem is that a single gene can be found under several different names, most often because different scientists working with different problems in different contexts have discovered it without knowledge about each other. These observations should be taken into account when designing new symbols and they also highlight the urgent need of a standardised human gene ontology. However, since there are already many ambiguous symbols used in the literature this problem with ambiguity will remain. Therefore, it is important to improve specificity of the indexing procedure, for instance, by incorporation of Natural Language Processing for term recognition, for example, as reported by Fukuda [2].

Preliminary investigations of the PubGene neighbours of the PPARA, PPARD, and PPARG revealed that many associations for these genes were missed due to incomplete lists of literature aliases (data not shown). Incomplete lists of literature aliases will naturally cause relevant associations to be missed. This problem can be reduced by manually editing the alias lists prior to indexing. This would require a lot of resources and, so far, nomenclature information for PubGene was collected from external databases and compiled without any editing. However, in order to distribute the cost of editing the alias information, it would be interesting to create an interactive user interface where it should be possible to add relevant aliases for a gene and then upgrade the network. Similarly, it should be possible to remove, or rather inactivate some synonyms that obviously cause problems such as false connections. Clearly, the 'NUC1' synonym of PPARD is a candidate for elimination.

Statistical analyses support our hypothesis that enzyme neighbours are more closely related than general pairs of enzymes. Three different, but related, analyses show that the distances between neighbours in PubGene are very different from what one would expect from a random sampling of enzyme pairs. The distribution of distances between enzyme pairs in PubGene is considerably skewed to the left (towards lower values) in Figure 2 compared to the background distribution. Consequently, the average distance is lower and the proportion of 0- or 1-distance pairs is higher. We interpret this as evidence that the PubGene network captures biologically relevant information.

It is clear that PubGene is not complete with respect to the definition of SCI neighbours. Since many of the missing SCI neighbours had relevant biological relationships, we see that the co-occurrence relationship is not capable of detecting all such relationships. A plausible reason is that in an abstract there is only a limited amount of space available and therefore, all interesting results cannot be mentioned there. This indicates that it may be beneficial to combine the information from SCI with that of PubGene to get a more biologically complete network of genes.

In this paper, we have used the definition of SCI neighbours to evaluate the PubGene network, but to create and evaluate a complete SCI network is an interesting direction for future work. However, it should be kept in mind that although we have sketched how we could create a complete SCI network, we have only examined two genes and their SCI neighbours. Clearly, more examples should be assessed before a general conclusion can be drawn. Furthermore, it should be noted that the definition of original articles excludes many of the genes in PubGene because not all genes can be related to a protein entry in SWISSPROT or TrEMBL, and also because not all such entries have literature references.

The ideas underlying the PubGene and SCI networks are related to those implemented in the ARROWSMITH system [12]. The ARROWSMITH system has been used to analyse article titles to relate what Swanson and Smalheiser call ‘complementary and noninteractive’ sets of articles. Their approach is to ‘discover’ implicit relationships between, for instance, a disease D, such as Alzheimer’s disease and an agent A, such as estrogen. Such an implicit relationship is said to be found if there exists a set of terms B such that A has been mentioned with one term in B, and D has been mentioned with one term in B, but A and D has not been mentioned together. Although our intention is not primarily to discover new relationships, but rather to extract those that already are known, the PubGene network as well as the SCI network contains associations that have not been explicitly mentioned by the authors. It would be interesting to explore how the PubGene network and the SCI approach could be used to create hypotheses about possible interactions.

A formal definition of correctness and completeness is difficult because there is no clear-cut definition of what is to be regarded as biologically relevant. It is obvious that many different types of relationships between two genes may exist, and biologically relevant in one context, is not necessarily biologically relevant in another. For instance, if regulatory interactions are of interest, it is not correct to associate two genes only because they code for proteins that belong to the same protein family. In fact, one type of relation may be very abundant, which may have the effect that other associations become hard to distinguish. As there are different notions of relevance, completeness at one level may only be attained if correctness at another level is given up. Ideally, one would have different networks reflecting different relationships. For example, one network that reflects functional similarity and a different network that reflects regulatory or metabolic pathways.

Even though our results demonstrate that improvements to PubGene would give a more correct and complete network, they also indicate that many biologically relevant associations have already been captured. By presenting a comprehensive summary of published knowledge, both explicitly and implicitly stated, the PubGene network may be used as a tool for generation of ideas and hypotheses. As such, PubGene is useful even if it is only partially correct and partially complete. Furthermore, our results have given a better understanding of what kinds of relationships are reflected by abstract co-occurrence. This way, our work contributes to a better interpretation of the information in PubGene, as well as information in similar databases.

Acknowledgements

We thank the National Library of Medicine for giving us access to the MEDLINE data. Also we would like to thank Petter Hallgren (RKS Data AB) for helping us with database issues. We are grateful to Eivind Hovig, Staal Vinterbo, and Astrid Læg Reid for helpful discussions. This work has been supported in part by a grant from the Norwegian Cancer Society and Norwegian University of Science and Technology. TKJ was supported in parts by grant 134422/410 from the Norwegian Research Council.

References

1. Bishop-Bailey, D., "Peroxisome proliferator-activated receptors in the cardiovascular system", *Brit. J. Pharmacol.*, **129**:823-834 (2000)
2. Fukuda, K. *et al*, "Towards information extraction: Identifying protein names from biological papers", in *Pacific Symposium on Biocomputing*, **3**:707-718 (1998)
3. He, T.-C. *et al*, "PPAR δ is an APC-regulated target of nonsteroidal anti-inflammatory drugs", *Cell*, **99**:335-345 (1999)
4. Jenssen, T-K. *et al*, "Pubgen: Discovering and visualising gene-gene relations", in *Currents in Computational Molecular Biology*, S. Miyano, R. Shamir, and T. Takagi, Editors, pp. 48-49 Tokyo (2000)
5. Jenssen, T-K. *et al*, "A literature network of human genes for high-throughput gene-expression analysis", *submitted for publication*
6. Klinge, C. M., "Estrogen receptor interaction with co-activators and co-repressors", *Steroids*, **65**:227-251 (2000)
7. Okamoto, K. *et al*, "Redox-dependent regulation of nuclear import of the glucocorticoid receptor", *J. Biol. Chem.*, **274**:10363-10371 (1999)
8. Palmer, C. N. A. *et al*, "Peroxisome proliferator activated receptor- α expression in human liver", *Mol. Pharmacol.*, **53**:14-22 (1998)
9. Rindflesch, T. C., Rajan, J. V., and Hunter, L., "Extracting molecular binding terms from biomedical text", in *Proceedings of the 6th Applied Natural Language Processing Conference*, pp 188-195 (2000)
10. Sekimizu, T., Park, H. S., and Tsujii, J., "Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts", in *Genome Informatics*, Universal Academy Press (1998)
11. Stapley, B. J. and Benoit, G., "Biobibliometrics: Information Retrieval and Visualization from Co-Occurrence of Gene Names in Medline Abstracts", in *Pacific Symposium on Biocomputing*, **5**:526-537 (2000)
12. Swanson, D. R. and Smalheiser, N. R., "An interactive system for finding complementary literatures: a stimulus to scientific discovery", *Artificial Intelligence*, **91**:183-203 (1997)
13. Thomas, J. *et al*, "Automatic Extraction of Protein Interactions from Scientific Abstracts", in *Pacific Symposium on Biocomputing*, **5**:538-549 (2000)
14. Tipton, K. and Boyce, S., "History of the enzyme nomenclature system", *Bioinformatics*, **16**:34-40 (2000)
15. Tripoidis, N. *et al*, "Construction of a high-resolution 2.5-Mb transcript map of the human 6p21.2-6p21.3 region immediately centromeric of the major histocompatibility complex", *Genome Research*, **10**:454-472 (2000)

16. White, J. A. *et al*, "Guidelines for human gene nomenclature (1997). HUGO Nomenclature Committee", *Genomics*, **45**:468-471 (1997)
17. Willson, T. M. *et al*, "The PPARs: From orphan receptors to drug discovery", *J. Med. Chem.*, **43**: 527-550 (2000)