A CitationRank algorithm inheriting Google technology designed to highlight genes responsible for serious adverse drug reaction

Lun Yang1,2,*, Langlai Xu3 and Lin He1,2,4,*

1Bio-X Center, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Shanghai Jiao Tong University, Shanghai 200030, 2Institutes of Biomedical Sciences, Fudan University, 138 Xueyuan Road, Shanghai, 200032, 3College of Life Sciences, Nanjing Agricultural University, 1 Weigang, Nanjing, 210095 and 4Institute for Nutritional Sciences, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai, 200031, PR China

Received on March 12, 2009; revised and accepted on June 9, 2009

ABSTRACT

Motivation: Serious adverse drug reaction (SADR) is an urgent, world-wide problem. In the absence of any well-organized gene-oriented SADR information pool, a database should be constructed. Since the importance of a gene to a particular SADR cannot simply be defined in terms of how frequently the two are cited together in the literature, an algorithm should be devised to sort genes according to their relevance to the SADR topics.

Results: The SADR-Gengle database, which is made up of gene-SADR relationships extracted from PubMed, has been constructed, covering six major SADRs, namely cholestasis, deafness, muscle toxicity, QT prolongation, Stevens–Johnson syndrome and torsades de points. The CitationRank algorithm, which inherits the principle of the Google PageRank algorithm that a gene should be highly ranked when biologically related to other highly ranked genes, is devised. The algorithm performs robustly in recovering SADR-related genes in the presence of extraneous noise, and the use of the algorithm has been extended to sorting genes in our database. Users can browse genes in a Google-type system where genes are ordered according to their descending relevance to the SADR topic selected by the user. The database also provides users with visualized gene-gene knowledge chain networks, helping them to systematize their gene-oriented knowledge chain whilst navigating these networks.


Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Serious adverse drug reaction (SADR) has always been a concern, particularly when Vioxx® events (Furberg, 2006) and Avandia® events (Nissen and Wolski, 2007) are reported. Better drug safety could be achieved if the genetic risk factors and the mechanisms of SADR (Wilke et al., 2007) could be identified. However, candidate gene selection is hampered by a lack of knowledge of the SADR mechanism (Need et al., 2005). Genes at the pharmacokinetic level, e.g. the drug metabolite enzymes or transporters, do not give a satisfactory explanation for type B SADRs, namely the dose-independent ADRs (Pirmohamed and Park, 2001). Genome-wide association studies are handicapped by the lack of case-control samples, and their results cannot be properly interpreted without knowledge of the molecular mechanism of SADR. Researchers therefore need a database pooling comprehensive information on SADR-associated genes. For example a researcher exploring the Stevens–Johnson syndrome (SJS) should be able to access the relevant genes and their relationships with SJS on a relevant semantic web. Such gene-oriented databases do exist, covering gene-disease (Allen et al., 2008; Bertram et al., 2007; Lin et al., 2006) and gene-drug (Altman, 2007) relationships. However, to our knowledge, no database has so far been set up to cover the gene-SADR relationship.

The construction of a gene-oriented SADR database could be accomplished by text-mining the SADR literature. However, one obstacle is the entry recognition of gene names (Jensen et al., 2006). In addition, it is not easy to identify from a PubMed entry whether the corresponding species is Homo sapiens. GeneRIF (Gebel et al., 2008) of Entrez Gene includes a semi-automatic index of gene-Pubmed relationships examined by database curators. Furthermore, a gene2pubmed index was constructed from the GeneRIF and the GenBank. It possesses high precision but low sensitivity in retrieving SADR-related genes. The major problem is that there is no guarantee that gene A is less important than gene B to the SADR X simply because it is less frequently cited in X-related literature than B. Google’s PageRank algorithm (Brin and Page, 1998) presumes that a web page should be highly ranked if other highly ranked pages have hyperlinks to it. Morrison et al. (2005) extended this major principle, and successfully applied it to prioritize genes from a noisy microarray dataset. The algorithm could be extended to enrich and sort SADR-related genes, so that a gene would be assigned a high relevance to X if it is co-cited (Jensen et al., 2001) with other genes of high relevance to X, even if it is less frequently mentioned in X-related literature. The PageRank algorithm adopts the random walk hypothesis on a probability model. A recent study interpreting information flow within a network utilized an information-theoretic approach on the random walk model (Rosvall and Bergstrom, 2008), and was further applied in measuring the impact of journals (Bergstrom et al., 2008). Based on these algorithmic advances, we put forward an algorithm implementing the random walk model.
to evaluate genes’ importance in the literature. The algorithm was set up upon a gene–gene knowledge chain network (GKCN) and its effectiveness has been tested on several noisy datasets. We have therefore utilized it to sort genes in our own database named SADR Cengage, with the objective of providing gene–SADR data on user-friendly interfaces.

2 METHODS

2.1 Construction of SADR-oriented bibliome and the genes’ knowledge chain network

The six SADRs included in the database were those usually reported in the literature, which were directly or indirectly associated with X, and the CRR of a gene set to zero. For X, the core and extended genes can be regarded as genes less than three would be regarded as extended genes, with their CRR being less than zero. Genes whose CRR were greater than zero with either of their neighbors. The gene–Pubmed index called gene2pubmed was downloaded from eSearch and eFetch APIs, and were deposited into a rational database (MySQL 6.0). The reference impact factor for each Pubmed entry is the mean reference impact factor between two co-cited genes in an entry, they tend to relate to each other biologically directly or indirectly (Hoffmann and Valencia, 2003, 2004; Jenssen et al., 2001). A connection between two genes is established if they are co-cited in at least two Pubmed entries. The GKCN of the six SADRs topics was constructed based on the genes’ co-citation, and were further transformed into XGMML (Shannon et al., 2003) so as to be visualized in the jPlead (Klammer et al., 2008) Applet.

2.2 Assignment of the core genes and the extended genes

We were able to retrieve genes indexed in SADR X-related Pubmed from the gene2pubmed index. For a gene i, the \( a_i, b_i, c_i, d_i \) values, representing the number of Pubmed entries containing gene (\( a_i \)), not containing gene (\( c_i \)) under X or non-X, respectively (Supplementary Table 1), were counted and the citation rate ratio (CRR) was calculated as:

\[
\text{CRR} = \frac{a_i}{a_i + b_i + c_i + d_i} \tag{1}
\]

A gene was assigned as a core gene if its CRR exceeded zero with a \( a_i \) exceeding three. This threshold was set to ensure the gene’s knowledge was fully disclosed in X and non-X. Genes whose CRR was equal to zero but were connected to the core gene in the GKCN were assigned as extended genes. Genes whose CRR were greater than zero with either of their \( a_i \) or \( b_i \) less than three would be regarded as extended genes, with their CRR being zero or less. Genes which were directed or indirectly associated with X, and the CRR of a gene to some extent reflected the importance of the gene in topic X.

2.3 The CitationRank algorithm

Sorting SADR-related genes by their CRR value potentially creates false negatives. For example, it would be wrong to assign a low importance to a gene because currently its citation rate is low, since our knowledge of the molecular mechanism of SADRs is limited, especially when the gene is biologically linked to other genes with high CRRs, it remains unclear, therefore, whether this gene should be omitted. This problem is particularly acute in SADR research, where knowledge of all SADR-related genes is scarce anyway. Following the logic of the Google PageRank algorithm (Brin and Page, 1998) and the Eigenfactor algorithm (Borgstrom et al., 2008), we put forward an algorithm named CitationRank. PageRank is based on the premise that the original rank of page, can be measured by the probability 1–0, which is the probability of an internet surfer randomly choosing it over all web pages. The surfer can also arrive at page, with probability \( d \) from other web pages holding hyperlinks to it. In CitationRank, the original rank of gene, is defined as the likelihood that a researcher would access it in the course of looking at papers of a specific SADR topic whilst browsing the literature. We use \( (1-d)\text{CRR} \) as a measure of this possibility. The researcher can also ‘think of’ a gene, when looking at other genes which are co-cited with gene, in the literature. The links in the algorithm can be defined as edges in GKCN. Thus in the following iteration equation, the citation rank of gene\( (e_i) \) consists of two terms,

\[
e_i^{(t)} = (1-d)\text{CRR}_i + d \sum_{j \in N_i} \frac{\text{CRR}_j}{\text{PageRank}_j}, 1 \leq i \leq N, i \neq t \tag{2}
\]

where \( e_i^{(t)} \) denotes the citation rank (CR) of gene i, in the tth iteration. The initial rank vector is taken as \( e^{(0)} = \text{CRR} / ||\text{CRR}|| \). \( W \in R^{n \times n} \) is the connectivity matrix of GKCN, with \( wij \) representing the number of Pubmed entries where gene i and gene j are co-cited. The lines, equal to \( \sum_j wij \), The CitationRank uses the parameter \( d \) in the range \([0, 1]\) to control the weighting of CR and the network connectivity in the rank calculations. It has been proved that convergence of the iteration is guaranteed for all \( 0 < d < 1 \) (Morrison et al., 2005). Hence the Jacobi iteration was performed to solve the vector \( R \).

2.4 Test for the robustness of CitationRank algorithm against false negative noises

The aim of applying the CitationRank is to solve the problem of the false negative while using CRR to prioritize SADR-related genes. Here we conducted the ‘leave one out cross validation’ (LOOCV) to test the effectiveness of the algorithm when noises of the false negatives were introduced. For SADR X, the CitationRank vector \( R \) for X-related genes, including core and extended genes, was first calculated. Genes were sorted by the ascending CR to generate an order vector \( R \). Then for M core genes, the observation of the CRR of gene, was set to zero in turn, and the vector \( R \) was re-calculated based on \( Q \). CR observations. So an order vector \( O_i \) was then constructed where the re-ordered position for the omitted gene, can be recovered as \( O_i \). Here the ‘boosting’ index \( B \) for gene i was computed as:

\[
B_i = \frac{O_i}{O_{\max}}, 1 \leq i \leq N \tag{3}
\]

and the goodness of recalling the genes towards their original position in the ranking list in the LOOCV was measured by the \( C \) value, denoting the percentage of genes whose index \( B \) were greater than 0.9.

2.5 Test for the robustness of the CR value-based binary classification model

A gene can be classified as a core or an extended gene when a threshold of a certain classification variable of the gene is set. Different ROC curves can be drawn using different classification variables, and the effectiveness of the classification can be measured by the area under curves (AUCs) of the ROC curves. The ‘1–specificity’ and ‘sensitivity’ of the x-axis and the y-axis are defined as:

\[
\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{5}
\]

where \( FP, TN, TP \) and \( FN \) denote the number of false positives, true negatives, true positives and false negatives. For SADR X, the boundary between core
and extended classes was determined by CRR and CR value, respectively. To test the robustness of the classifier against false negative noise, we randomly assigned the CRR values of N core genes to zero. The strength of the noise was controlled by N. Different ROC curves were drawn by using different classification variables with different N values and stepwise d values.

2.6 Enrichment of SADR associated pathways

For SADR X, core and extended genes with high CR value greater than 0.001 were included in the enrichment analysis of KEGG Orthology (Kanehisa et al., 2008) using KOBAS (Mao et al., 2005). A threshold at 0.05 of the q value (Strimmer, 2008) was used to choose the specifically enriched pathways. The enriched terms and the contributing genes were deposited in our database.

3 RESULTS

3.1 Robustness of the CitationRank algorithm against false negative noises

We took the muscle toxicity related gene set as an example. We set the CRR of the core genes in turn to zero to intentionally generate false negatives. The CR value responded robustly to this anomaly, as most of the wrongly assigned core genes were recovered close to their original place in the ranked list (Fig. 1b–d). The performances of LOOCV were measured by C values. The best C was 0.63, which was achieved when the parameter d was set to 0.9 (Fig. 1d), namely 74 of 118 core genes being recalled in 90% of their original positions. However, a portion of genes failed to be recovered by CitationRank irrespective of what d values were set. This was due to their lack of connection in the GKCN. These ‘orphan’ genes could not share the rank value spread from other genes. Similar trends of the algorithm could be observed in other SADR-related gene sets (Supplementary Figs 1 and 2).

3.2 The robustness of the CR value based classification model

We further tested the power of mapping muscle toxicity related genes onto core and extended classes using CRR and CR as the classification variable. Given that the total number of the core genes was 118, parameter N was set at 30, 60 and 90 in three tests. According to the Equation (2), the CR vector was equivalent to the CRR vector when d equals zero. The lowest AUC was always observed when CRR was used as the classification variable (d=0 in Fig. 2a) whichever N was applied. A higher d-value resulted in higher AUCs (Fig. 2). When the noise was not strong, namely N was set at 30 and 60 (Fig. 2a and b), the CR-based classifier did not perform significantly better than the CRR-based classifier, and the d parameter did not seem to be a key factor in determining the AUC. However, when N was set at 90 (Fig. 2c), the AUC decreased significantly when d was set at 0 and 0.5 compared to the corresponding results in Figure 2a and b, but the AUC remained unchanged in the classifier when d was set at 0.9. Even when these 90 core genes (76%) were wrongly assigned, an AUC of 0.81 (Fig. 2c) was still achieved with this classifier, indicating that considerable...
Fig. 2. ROC curves representing the power of classifying core genes and extended genes using CRR and CR respectively. Here \( N \) denotes the number of wrongly assigned core genes. The CRR vector was equivalent to the CR vector when \( d \) was set at zero. (a) The AUC was 0.87, 0.90 and 0.89 when \( d \) was set at 0, 0.5 and 0.9, respectively. (b) The AUC was 0.75, 0.81 and 0.81 when \( d \) was set at 0, 0.5 and 0.9, respectively. (c) The AUC was 0.62, 0.67 and 0.81 when \( d \) was set at 0, 0.5 and 0.9, respectively.

Table 1. Key statistics of SADR Gengle

<table>
<thead>
<tr>
<th>Topic</th>
<th>Pubmed retrieved</th>
<th>Pubmed indexed with gene</th>
<th>Number of core gene</th>
<th>Number of extended genes</th>
<th>Number of pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholestasis</td>
<td>28296</td>
<td>258</td>
<td>18</td>
<td>3629</td>
<td>74</td>
</tr>
<tr>
<td>Deafness</td>
<td>53099</td>
<td>1385</td>
<td>59</td>
<td>4659</td>
<td>66</td>
</tr>
<tr>
<td>Muscle Toxicity</td>
<td>119193</td>
<td>1828</td>
<td>118</td>
<td>9474</td>
<td>65</td>
</tr>
<tr>
<td>QT Prolongation</td>
<td>4931</td>
<td>336</td>
<td>10</td>
<td>2764</td>
<td>56</td>
</tr>
<tr>
<td>Stevens–Johnson Syndrome</td>
<td>18566</td>
<td>101</td>
<td>7</td>
<td>651</td>
<td>64</td>
</tr>
<tr>
<td>Torsades de Pointis</td>
<td>2128</td>
<td>32</td>
<td>3</td>
<td>283</td>
<td>10</td>
</tr>
</tbody>
</table>

All statistics were made on 1 March 2009.

3.3 Browsing the SADR-related genes in SADR Gengle

As described above, the CitationRank algorithm can solve the problem of the false negatives, indicating that the basic premise of the algorithm is sound. This premise was then naturally extended to the sorting of the core genes in our database. Genes should be sorted and presented to the client by their decreasing relevance to topic X, and such relevance was measured by the CR values in our database. In Google, web pages relevant to the user’s key word are sorted mainly based on their PageRank (Brin and Page, 1998). In our system one can browse the SADR-related genes as if browsing the search results from Google. The key data statistics of SADR Gengle was listed in Table 1. Taking SJS for example, core genes are sorted by their descending CR and are presented as gene cards with general information presented to help the user to pursue their interests. Within each card, the relevant extended genes, if any, in GKCNI are displayed, and the SJS-related literature containing this gene is displayed. Here the impact factor of the gene denotes the mean reference impact factor of the SJS-related Pubmed entries carrying this gene. One can go to the detailed information page by clicking on the gene name. In the detailed page of HLA-B under SJS, all SJS-related papers describing HLA-B are listed. Gene ontology terms, KEGG pathways and OMIM entries of HLA-B, if any, are displayed. Detailed information of any ‘child’ genes can be retrieved by navigating the ‘NETWORK NAVIGATOR’ on the right of the page.

3.4 Case study of the CitationRank algorithm

To comprehend the parameter \( d \) in the Equation (2) and how could robustness be achieved against noises, we examined the CR values’ change within a local network of ‘muscle toxicity’ related genes by assigning different \( d \) values (Fig. 3). All genes were extended genes

---

A CitationRank algorithm inheriting Google technology
The GKCN was used to define the linkages between all genes TOR1A, a gene from the Applet through right clicking on a particular gene in blue dashed rectangle. One can also access the detailed page of genes are in grey dashed rectangles, whereas extended gene knowledge chain in the visualized GKCN (part V in Fig. 4). Core knowledge chain of SJS. Furthermore, one can also identify this core and extended genes in such a way to retrieve a gene-oriented knowledge is presented (part IV in Fig. 4). Users can jump between another navigator (part III in Fig. 4) where the 'parent' genes of following the link of an extended gene (in GKCN, the extended genes, are listed in the network navigator. Pubmed entries carrying to the detailed information page (part II in Fig. 4), where SJS-related genes are sorted by their CR values, and are presented in a Google pattern. Clicking on HLA-B will lead the user to the detailed information page (part II in Fig. 4), where SJS-related Pubmed entries carrying HLA-B will be displayed. Its neighbors in GKCN, the extended genes, are listed in the network navigator. Following the link of an extended gene (LTA) will lead the user to another navigator (part III in Fig. 4) where the 'parent' genes of LTA are displayed. Clicking on IL4R in the navigator, for example, will take the user to another detailed page, where some new gene-SJS knowledge is presented (part IV in Fig. 4). Users can jump between core and extended genes in such a way to retrieve a gene-oriented knowledge chain of SJS. Furthermore, one can also identify this knowledge chain in the visualized GKCN (part V in Fig. 4). Core genes are in grey dashed rectangles, whereas extended gene LTA is in blue dashed rectangle. One can also access the detailed page of a gene from the Applet through right clicking on a particular gene to retrieve a popup link. In this example, the user has constructed a knowledge chain of 'SJS–HLA-B–LTA–IL4R–SJS' extracted from GKCN. He could deduce that LTA relates to HLA-B and IL4R, two core genes that are involved in the pathogenesis of SJS (Hung et al., 2005; Ueta et al., 2007), implying a putative functional linkage of this extended gene to SJS through MHC I-mediated pathways (Chesman et al., 2008) or cytokine-mediated pathways (Ueta et al., 2007). On the other hand, the logical connectivity of these three genes might prompt the user to conceive the hypothesis that a crosstalk between a MHC I-related pathway and a cytokine-related pathway might be mediated by LTA, which might give a systematic explanation of the pathogenesis of SJS.

Another case of using GKCN to retrieve potential molecular knowledge about SADR again concerns TOR1A, a muscle disease associated gene (Zorzi et al., 2009), which co-cited with 'muscle toxicity' associated literature (Wong et al., 2005) but does not meet the criteria for becoming a core gene. However, it links to the core gene, SGCE, which plays an important role in myoclonus dystonia (Ettinger et al., 1997). By navigating network navigator or the GKCN, the user can easily identify TOR1A as a potential muscle toxicity related gene, for it not only achieved a CR of 0.11, the seventh highest CR among 1763 extended genes, but was also identifiable as a 'child' gene of SGCE in the visualized GKCN. The examples above indicate that the SADR Gengle could help to mine embedded knowledge, generating important hypotheses for prospective validation of the candidate genes for this poorly understood subject.

SADR Gengle focuses on providing the bibliomic information (Sealls, 2001), the gene sorting algorithm and the methodology for user to navigate the literature. Users might also refer to protein-protein interaction (PPI) and gene co-expression data while reading the literature in the database. To facilitate the users, we provide hyperlinks for each gene to the STRING server (Jensen et al., 2009), which harbors comprehensive PPI and co-expression data.

3.6 Enrichment of SADR associated pathways
To achieve a more precise enrichment of pathways that were shared specifically by the genes under a certain SADR topic, we included core genes and highly ranked extended genes for enrichment analysis. Pathways enriched were usually consistent with existing except for SGCE and HLA-DRB1, whose CR values were equivalent to the CRR values when d was zero (Fig 3a). The size of each node reflected their CR values. After setting d at 0.5, the CR value of SGCE spread to LAMA1, TOR1A and HLA-DRB1 (Fig 3b). Such ‘spread effect’ became stronger when d was set at 0.9 (Fig. 3c).

Here are two ways to navigate the GKCN. Taking SJS for example again, one can jump from one gene to its ‘neighbors’ using network navigator or the visualized GKCN in the Applet. The navigation can begin from the gene list page in part I of Figure 4, where SJS-related genes are sorted by their CR values, and are presented in a Google pattern. Clicking on HLA-B will lead the user to the detailed information page (part II in Fig. 4), where SJS-related Pubmed entries carrying HLA-B will be displayed. Its neighbors in GKCN, the extended genes, are listed in the network navigator. Following the link of an extended gene (LTA) will lead the user to another navigator (part III in Fig. 4) where the ‘parent’ genes of LTA are displayed. Clicking on IL4R in the navigator, for example, will take the user to another detailed page, where some new gene-SJS knowledge is presented (part IV in Fig. 4). Users can jump between core and extended genes in such a way to retrieve a gene-oriented knowledge chain of SJS. Furthermore, one can also identify this knowledge chain in the visualized GKCN (part V in Fig. 4). Core genes are in grey dashed rectangles, whereas extended gene LTA is in blue dashed rectangle. One can also access the detailed page of a gene from the Applet through right clicking on a particular gene to retrieve a popup link. In this example, the user has constructed a knowledge chain of ‘SJS–HLA-B–LTA–IL4R–SJS’ extracted from GKCN. He could deduce that LTA relates to HLA-B and IL4R, two core genes that are involved in the pathogenesis of SJS (Hung et al., 2005; Ueta et al., 2007), implying a putative functional linkage of this extended gene to SJS through MHC I-mediated pathways (Chesman et al., 2008) or cytokine-mediated pathways (Ueta et al., 2007). On the other hand, the logical connectivity of these three genes might prompt the user to conceive the hypothesis that a crosstalk between a MHC I-related pathway and a cytokine-related pathway might be mediated by LTA, which might give a systematic explanation of the pathogenesis of SJS.

Another case of using GKCN to retrieve potential molecular knowledge about SADR again concerns TOR1A, a muscle disease associated gene (Zorzi et al., 2009), which co-cited with ‘muscle toxicity’ associated literature (Wong et al., 2005) but does not meet the criteria for becoming a core gene. However, it links to the core gene, SGCE, which plays an important role in myoclonus dystonia (Ettinger et al., 1997). By navigating network navigator or the GKCN, the user can easily identify TOR1A as a potential muscle toxicity related gene, for it not only achieved a CR of 0.11, the seventh highest CR among 1763 extended genes, but was also identifiable as a ‘child’ gene of SGCE in the visualized GKCN. The examples above indicate that the SADR Gengle could help to mine embedded knowledge, generating important hypotheses for prospective validation of the candidate genes for this poorly understood subject.

SADR Gengle focuses on providing the bibliomic information (Sealls, 2001), the gene sorting algorithm and the methodology for user to navigate the literature. Users might also refer to protein-protein interaction (PPI) and gene co-expression data while reading the literature in the database. To facilitate the users, we provide hyperlinks for each gene to the STRING server (Jensen et al., 2009), which harbors comprehensive PPI and co-expression data.
A CitationRank algorithm inheriting Google technology

Fig 4. Navigating the SJS-oriented genes’ co-citation network in SADR Gengle using ‘NETWORK NAVIGATOR’ and visualized GKCN. (I) The gene list page of the SJS topic in a ‘Google style’. Here genes are ordered by their CR value calculated from the CitationRank algorithm. (II) Detailed information page of HLA-B. The Pubmed entry highlighted describes the relationship between HLA-B and SJS. (III) Network navigator page of the extended gene LTA. Users can jump back to the detailed page of HLA-B, its parent gene, or jump to a new detailed page of another parent gene, e.g. IL4R. (IV) Detailed information page of IL4R. The Pubmed entry highlighted might describe another pathogenesis pathway of SJS other than the MHC I-mediated pathway, namely the cytokine-mediated pathway. (V) Knowledge chain of ‘SJS–HLA-B–LTA–IL4R–SJS’ in the GKCN visualization page. Note that some of the original screen shot was discarded for the sake of brevity.

knowledge. In the case of the SJS topic, for example, ‘T cell receptor-signaling pathway’ \( (q = 9.8E–20) \), ‘Antigen processing and presentation’ \( (q = 1.4E–17) \) and ‘B cell receptor signaling pathway’ \( (q = 1.0E–8) \) tended to occur more frequently in SJS-related genes than in a random human gene selection at the significance level of 0.01. The result corresponded with the known SJS mechanism (Borchers et al., 2008), indicating that the genes involved in the analysis, which was highlighted by CitationRank, could to some extent represent the molecule mechanism of this SADR. Furthermore, most of the results could be replicated by another enrichment tool (DAVID) (Huang da et al., 2007). When the enrichment results of six SADRs were summarized (see SADR
Geng online), most of them were found to share several common pathways, such as ‘Gap junction’, ‘Toll-like receptor signaling pathway’ and ‘Adherens junction’. Although these SADRs are triggered by different drugs and occurred in different tissue, the sharing of the same pathways implied that these common biological procedures might be responsible for the pathogenesis of the SADR and thus were worthy of experimental validation.

4 CONCLUSIONS

(1) The CitationRank algorithm has the potential for sorting genes by their relevance to a topic, and is effective in uncovering false negatives of the gene set relevant to a particular topic;
(2) At a time when molecular mechanisms are poorly understood and gene-oriented knowledge is not well organized, SADR Gengle enables users to broaden and systematized their gene-oriented knowledge on SADR;
(3) The SADR Gengle is an instance of the Gengle knowledge depositing model, which inherits some of the ‘genes’ from Google, and is designed to organize structured gene-topic knowledge. The model could be instantiated onto topics other than disease, such as longevity, pain and reproduction etc., about which knowledge has began to accumulate but is poorly organized. Such information could be quickly organized on a gene-oriented basis using the methodology described in this study.

ACKNOWLEDGEMENTS

We appreciate Jian Chen, Shuing Huang, Zichun Hua, Bin Wang, Gongli Xia, Xiangzhe Zhang and Zhenhua Xia for helpful discussions. We are grateful to the developers of the Pubmed and Entrez Gene. We thank the suggestions from the anonymous reviewers in improving the manuscript.

Funding: China Postdoctoral Science Foundation (PSF) [20070420660 to L.Y.], Shanghai PSF [61444 to L.Y.], 2007CB947300] and the Shanghai Leading Academic Discipline Project [B205].

Conflict of Interest: none declared.

REFERENCES