AUTOMATICALLY GENERATING GENE SUMMARIES FROM BIOMEDICAL LITERATURE*

XU LING, JING JIANG, XIN HE, QIAOZHU MEI
CHENGXIANG ZHAI, BRUCE SCHATZ

Department of Computer Science and Institute for Genomic Biology
University of Illinois at Urbana-Champaign Urbana, IL 61801
E-mail: {xuling,jiang4,xinhe2,qmei2,czhai,schatz}@uiuc.edu

Biologists often need to find information about genes whose function is not described in the genome databases. Currently they must try to search disparate biomedical literature to locate relevant articles, and spend considerable efforts reading the retrieved articles in order to locate the most relevant knowledge about the gene. We describe our software, the first that automatically generates gene summaries from biomedical literature. We present a two-stage summarization method, which involves first retrieving relevant articles and then extracting the most informative sentences from the retrieved articles to generate a structured gene summary. The generated summary explicitly covers multiple aspects of a gene, such as the sequence information, mutant phenotypes, and molecular interaction with other genes. We propose several heuristic approaches to improve the accuracy in both stages. The proposed methods are evaluated using 10 randomly chosen genes from FlyBase and a subset of Medline abstracts about Drosophila. The results show that the precision of the top selected sentences in the 6 aspects is typically about 50-70%, and the generated summaries are quite informative, indicating that our approaches are effective in automatically summarizing literature information about genes. The generated summaries not only are directly useful to biologists but also serve as useful entry points to enable them to quickly digest the retrieved literature articles.

1. Introduction

The rise of modern genomics in the 21st century is catalyzing the necessity for gene annotation of new organisms, which are not model genetic organisms and whose gene functions are largely unknown. There are already an order of magnitude more organisms whose sequences are known

*This work is in part supported by the National Science Foundation under award numbers 0425852 and 0428472.

than those whose genetics is known, and the number of such new organisms is growing rapidly. As part of the BeeSpace project at the University of Illinois (www.beespace.uiuc.edu), we are developing fully automatic annotation methods for model organisms beyond the genetic models, using computational methods. In particular, we are annotating genome data about the honey bee Apis mellifera using new text processing technologies on biomedical literature combined with existing model genetic databases, especially about the fruit fly Drosophila melanogaster. This paper describes a component software that supports automatic summarization of gene descriptions from biomedical literature.

The generated summary covers six aspects of a gene: (1) Gene products; (2) Expression location; (3) Sequence information; (4) Wild-type function and phenotypic information; (5) Mutant phenotype; and (6) Genetical interaction. Such a summary not only is itself very useful, but also can serve as useful entry points to the literature through linking each aspect to the supporting evidence in the literature, allowing biologists to more easily keep track of new discoveries occurring in the literature. If gene summaries can be automatically generated with decent accuracy, we would be able to curate the databases for other model organisms equivalently well as FlyBase[7] did, but much more efficiently.

To the best of our knowledge, this is the first attempt to automatically generate such a structured summary of a gene from biomedical literature. We present a two-step method, retrieving relevant articles then extracting informative sentences from these articles for each aspect. In the retrieval step, we propose several heuristics to address gene name variations to improve the retrieval accuracy. In the extraction step, we exploit training sentences in existing curated databases and score a sentence for each aspect based on its content, location, and the document containing the sentence.

We evaluate the proposed method using 10 randomly chosen genes from FlyBase and a subset of Medline abstracts about Drosophila. The precision of the top selected sentences in the 6 aspects is about $50 - 70\%$ and the generated summaries are quite informative, indicating that our approaches are effective in automatically summarizing literature about genes. Since our method is quite general, it is likely to work on other organisms as well.

2. Related Work

Most existing studies of biomedical literature mining focus on automated information extraction, using natural language processing techniques to identify relevant phrases and relations in text, such as protein-protein interactions[1] (see [2,3] for reviews of these works). The information we extract is at the sentence level, which allows us to cover many different aspects of a gene and extract information in a more robust manner.

A problem closely related to ours was addressed in the Genomics Track in the Text REtrieval Conference (TREC) 2003, where the task was to generate descriptions about genes from Medline records. The major differences between this task and ours are: (1) The generated descriptions do not organize the information into clearly defined aspects. In contrast, we define six reasonable aspects of genes and propose new methods for selecting sentences for specific aspects. (2) In genomics track, the existing GeneRIF in LocusLink (http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene) can be used as training data, which makes the problem easier, while we are dealing with situations where no such resource is available.

Automatic text summarization, notably news summarization has also been extensively studied. According to the scheme given in a detailed review[4], our gene summarization task is a type of informative, query-oriented, multi-document extraction. Again, a distinctive feature of our work is that the generated summary has explicitly defined semantic aspects, whereas most news summaries are simply a list of extracted sentences. Despite this difference, our two-step process of generating a summary and some of our heuristics used in sentence selection are similar to what has been used for news summarization[5].

3. Automatic Gene Summarization

3.1. *Overview*

Our automatic gene summarization system mainly consists of two components: a Keyword Retrieval module that retrieves documents about a target gene, and an Information Extraction module that extracts sentences from the retrieved documents to summarize the target gene. The Information Extraction module itself consists of two components, one for training data generation, and the other for sentence extraction. The whole system is illustrated in Figure 1.

3.2. *Keyword Retrieval Module*

First, to identify documents that may contain useful information for the target gene, we use a dictionary-based keyword retrieval approach to retrieve all documents containing any synonyms of the target gene.

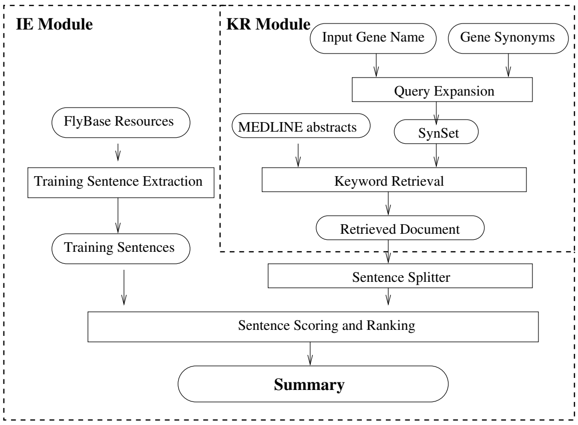

Figure 1. System Overview.

3.2.1. *Gene* SynSet *Construction*

Gene synonyms are very common in biomedical literature. It is important to consider all the synonyms of a target gene when searching for relevant documents about the gene. We used the synonym list for fly genes provided by BioCreAtIvE Task 1B[6] and extended it by adding names or functional information of proteins encoded by each gene from FlyBase's annotation. In the end, we constructed a set of synonyms and protein names (called *SynSet* here) for each known Drosophila gene.

To further improve the recall of retrieval, we investigated variations in gene name spelling. The following variations are identified and addressed in our system: (1) There are various ways to separate name constituents: they can be contiguous or separated by various separators such as white spaces, hyphens, slashes and brackets. (2) Gene names can be spelled in upper or lower case. To deal with these variations, our system uses a special tokenizer for both Medline abstracts and *SynSet* entries. The tokenizer converts the input text into a sequence of tokens, where each token is either a sequence of lowercase letters or a sequence of numbers. White spaces and all other symbols are treated as token delimiters. For instance, the different synonyms for gene *cAMP dependent protein kinase 2*, "PKA C2", "Pka C2", and "Pka-C2", are all normalized to the same token sequence "pka c 2" to allow them to match each other. A Medline abstract is considered as being relevant only if it matches the token sequence of a synonym *exactly*.

3.2.2. *Synonym Filtering*

Some gene synonyms are ambiguous, for example, the gene name "PKA" is also a chemical term with a different meaning. In these situations, a document containing the synonym with an alternative meaning would be retrieved. Our strategy of alleviating this problem is based on the observations that (1) the longer or full name of a gene is often unambiguous; (2) when a gene's short abbreviation is mentioned in a document, its full or longer name is often present as well. Therefore, we force all retrieved documents to contain at least one synonym of the target gene that is at least 5-character long.

3.3. *Information Extraction Module*

The information extraction module extracts sentences containing useful factual information about the target gene from the documents returned by the keyword retrieval module. To ensure the precision of extraction, we only consider sentences containing the target gene, which are further organized into the six general categories listed in Table 1, which we believe are important for gene summaries.

Table 1. Categories for Gene Summary

GP	Gene Product, describing the product (protein, rRNA, *etc.*) of the target gene.
EL	Expression Location, describing where the target gene is mainly expressed.
SI	Sequence Information, describing the sequence information of the target gene and its product.
WFPI	Wild-type Function & Phenotypic Information, describing the wild-type functions and the phenotypic information about the target gene and its product.
MP	Mutant Phenotype, describing the information about the mutant phenotypes of the the target gene.
GI	Genetical Interaction, describing the genetical interactions of the target gene with other molecules.

3.3.1. *Training Data Generation*

To help identify informative sentences related to each category, we construct a training data set consisting of "typical" sentences for describing each of the six categories using three resources: the *Summary* pages, the *Attributed data* pages, and the *references* of each gene in FlyBase.

The "Summary" Paragraph: FlyBase curators have compressed all the relevant information about a gene into a short paragraph, the text *Summary* in the FlyBase report. This paragraph contains good example sentences for each aspect of a gene. A typical paragraph contains information related to gene product, sequence information, genetical interaction, *etc.* More importantly, verbs such as "encode", "sequence" and "interact" in the text are very indicative of which category the sentence is related to. Based on the regular structure of these text summaries, we decompose each paragraph into our six categories with non-relevant sentences discarded.

However, since these sentences are generated from a common template by a curator, they are not good examples of typical sentences that appear in real literature. For instance, genetical interaction can be described in many different ways using verbs such as "regulate", "inhibit", "promote" and "enhance". In the "summary" paragraph, it is always described using the template "It interacts genetically with ...". Thus we also want to obtain good examples of original sentences from the literature.

The "Attributed Data" Report: One resource of original sentences is the "attributed data" report for each Drosophila gene provided by FlyBase. For some attributes such as "molecular data", "phenotypic info." and"wild-type function", the original sentences from literature are listed. These sentences seem to be good complements of the training data from the "summary" paragraph. In our system, we collect the sentences from "phenotypic info." and "wild-type function" as training sentences for the category *WFPI*.

The References: For categories such as "gene product" and "interacts genetically with", the "attributed data" reports only list the noun phrases related to the target gene, but do not show any complete sentences. In order to find the patterns of sentences containing such information, we exploit the links to the corresponding references given in the "attributed data" reports to find the PubMed ID of the reference. We then look for occurrences of the item, *i.e.,* a protein name in "gene product" or another gene name "interacts genetically with", in the abstract of the reference. We add the sentence containing both the item and the target gene to our training data. Inclusion of these sentences is useful because verbs such as "enhance" and "suppress" now appear in the training data.

3.3.2. *Sentence Extraction*

To extract sentences related to each category for a target gene, we first preprocess sentences by removing the stop words and stemming with a Porter stemmer. We then score each sentence as follows.

Category Relevance Score (S_c): We use the vector space model and cosine similarity function from information retrieval to assign a relevance score to each sentence *w.r.t.* each category. Specifically, For each category, we construct a corresponding term vector V_c using the training sentences for the category. Following a commonly used information retrieval heuristic, we define the weight of a term t_i in the category term vector for category j as $w_{i,j} = \mathrm{TF}_{i,j} * \mathrm{IDF}_i$, where $\mathrm{TF}_{i,j}$ is the term frequency, i.e., the number of times term t_i occurs in all the training sentences of category j, and IDF_i is the inverse document frequency. IDF_i is computed as $\mathrm{IDF}_i = 1 + \log \frac{N}{n_i}$, where N is the total number of documents in our document collection, and n_i is the number of documents containing term t_i. Intuitively, V_c reflects the usage of different words in sentences describing a category.

Similarly, for each sentence we can construct a sentence term vector V_s, with the same IDF and the TF being the number of times a term occurs in the sentence. The category relevance score is then the cosine of the angle between the category term vector and the sentence term vector: $S_c = \cos(V_c, V_s)$.

Document Relevance Score (S_d): A good sentence to be included in our summary should be both relevant to a category and informative. To measure the informativeness of a sentence, we compute a document relevance score for each sentence, which is the cosine similarity between the sentence vector V_s and the document vector V_d, which is computed similarly to the other vectors described above.

Location Score (S_l): A useful heuristic for news article summarization is to favor sentences at the beginning of a document. For scientific literature, however, the last sentence of an abstract is usually a summary of the experimental results or the discovery. Therefore, we also assign each sentence a location score, which is 1 for the last sentence of an abstract, and 0 otherwise.

Sentence Ranking and Summary Generation: The final score of a sentence S is a weighted sum of the three scores mentioned above with the weights set empirically: $S = 0.5S_c + 0.3S_d + 0.2S_l$. To ensure reliable association between sentences and categories, for each sentence, we rank all the categories based on S and keep only the top two categories. To generate a structured, category-based summary, for each category, we rank all the kept sentences according to S and pick the top-k sentences. Such a category-based summary is similar to the “attributed data” report in FlyBase. We also generate a paragraph-long summary by combining the top sentences of all the categories in the following way: We “grow” our paragraph summary by taking a top-ranked sentence from each category that is different from all the already included sentences in the paragraph summary. We impose an order on the categories so that the most specialized category would have a chance to contribute a sentence first.

4. Experiments and Evaluation

4.1. *Experiment Setup*

We retrieved 22092 Medline abstracts as our document collection using the keyword "Drosophila". We used the Lemur Toolkit[a] to implement the keyword retrieval module. Lemur is a C++ toolkit supporting a variety of information retrieval functions. We mainly exploited its indexing capability to quickly retrieve documents containing a given keyword. Our gene summarization algorithm runs very fast, taking only seconds to generate a summary on a Dell PowerEdge 2650 (3.06GHz CPU, 4GB Memory).

We used about 1/5 of the training data in FlyBase for training and randomly selected 10 genes from FlyBase for evaluation. For each gene, we ran three experiments. The first is a baseline run (*BL*), in which we randomly select k sentences. In the second run (*CatRel*), we use Category Relevance Score S_c to rank sentences. In the third run (*Comb*), we use combined score S to rank sentences.

4.2. *Evaluation and Discussion*

For each category of each gene, we generated top-k sentences from each run, and then asked two annotators with domain knowledge to judge the relevance. A sentence is considered to be relevant to a category if and only if it contains information on this aspect, regardless whether it contains any extra information. The evaluation metric is the precision of the top-k sentences for each category. The results are shown in Table 2. The average precisions of top-10 sentences for most categories by the two ranking methods are about $50 - 70\%$, while the average precision by random selection is typically about 20%. In most cases, combining all three scores performs only slightly better than using the Category Relevant Score alone. This could either be due to the fact that we use a simple function to combine the three scores and the parameters are not fully optimized, or suggest that those general text summarization heuristics may not be applicable to our problem.

[a] http://www.lemurproject.org/

We notice that the improvements over the baseline are most pronounced for categories *EL*, *SI*, *MP* and *GI*. This may be because these four categories are more specific and thus harder to detect by random selection.

Table 2. Precision of the top-k extracted sentences

		Avg. Precision					Avg. Precision		
cat.	top-k	*BL*	*CatRel*	*Comb*	cat.	top-k	*BL*	*CatRel*	*Comb*
EL	1	0.1	**0.9**	0.85	MP	1	0.1	**0.6**	0.55
	2	0.1	**0.8**	0.73		2	0.13	0.53	**0.55**
	5	0.14	**0.58**	0.58		5	0.13	0.36	**0.43**
	10	0.18	0.48	**0.51**		10	0.17	0.33	**0.45**
GP	1	0.45	**0.8**	0.75	GI	1	0.1	**0.7**	0.7
	2	0.43	0.78	**0.8**		2	0.13	**0.68**	0.65
	5	0.42	0.73	**0.75**		5	0.21	0.62	**0.67**
	10	0.4	0.57	**0.67**		10	0.23	0.56	**0.58**
SI	1	0.1	**0.85**	0.85	WFPI	1	0.45	**0.6**	0.55
	2	0.05	0.78	**0.8**		2	0.58	**0.78**	0.73
	5	0.12	0.63	**0.66**		5	0.6	**0.78**	0.77
	10	0.15	0.49	**0.54**		10	0.6	0.73	**0.75**

In Table 3, we show a sample structured summary generated for the well-studied gene *Abl*, in which all the extracted sentences are quite informative as judged by biologists. For comparison, we show the human-generated FlyBase summary of the same gene in Table 4.

To see how well our system performs on a less-studied gene, we show a sample structured summary generated for the less-studied gene *Camo*\Sod in Table 5. In this case, some sentences are not very relevant. However, by reading this summary, a biologist could still get some basic idea of the gene *Camo\Sod*. We cite one possible reconstruction of information based solely on our results in Table 5:

> *Camo*\Sod encodes the protein, CuZn superoxide dismutase, involved in superoxide production. In Drosophila, it is suggested that this gene is expressed in central nervous system. All the protein's important amino acids are conserved in related organisms. The mutation of this gene is known to be lethal.

The FlyBase summary for this gene is shown in Table 6, which is seen to be very short and barely informative. Considering that we have used no external information, the rich information content of our results is a strong indication of the usefulness of our system.

One problem of predefined categories is that not all genes fit into this framework. For instance, the gene *Amy-d* is an enzyme involved in carbohydrate metabolism and not typically studied by genetic means. As a

Table 3. Text summary of gene *Abl* by our system

GP	The Drosophila melanogaster abl and the murine v-abl genes encode tyrosine protein kinases (TPKs) whose amino acid sequences are highly conserved.
EL	In later larval and pupal stages, abl protein levels are also highest in differentiating muscle and neural tissue including the photoreceptor cells of the eye. abl protein is localized subcellularly to the axons of the central nervous system, the embryonic somatic muscle attachment sites and the apical cell junctions of the imaginal disk epithelium.
SI	The DNA sequence encodes a protein of 1520 amino acids with sequence homology to the human c-abl proto-oncogene product, beginning at the amino terminus and extending 656 amino acids through the region essential for tyrosine kinase activity.
MP	The mutations are recessive embryonic lethal mutations but act as dominant mutations to compensate for the neural defects of abl mutants.
GI	Mutations in the Abelson tyrosine kinase gene show dominant interactions with fasII mutations, suggesting that Abl and Fas II function in a signaling pathway that controls proneural gene expression.
WFPI	We have examined the expression of the abl protein throughout embryonic and pupal development and analyzed mutant phenotypes in some of the tissues expressing abl. abl protein, present in all cells of the early embryo as the product of maternally contributed mRNA, transiently localizes to the region below the plasma membrane cleavage furrows as cellularization initiates.

Table 4. Text summary of gene *Abl* from FlyBase

D. melanogaster gene ***Abl tyrosine kinase***, abbreviated as ***Abl***, is reported here. It has also been known in FlyBase as CG4032 and l(3)04674. It encodes a product with protein-tyrosine kinase activity (EC:2.7.1.112) involved in axon guidance which is localized to the axon; it is expressed in the embryo (embryonic central nervous system) and ovary (oocyte and ovary). It has been sequenced and its amino acid sequence contains a protein kinase, a SH2 motif, a tyrosine protein kinase, a SH3, a tyrosine protein kinase, active site and a protein kinase-like. It has been mapped cytologically to 73B1–4. It interacts genetically with Nrt, ena, fax, Lar, robo and 17 other listed genes. There are 28 recorded alleles: 15 in vitro constructs (none available from the public stock centers), 12 classical mutants (3 available from the public stock centers) and 1 wild-type. Amorphic mutations have been isolated which affect the central nervous system, the longitudinal connective, the commissure and 5 other listed tissues and are pupal recessive lethal, reduced (with Df(3L)st-j7) viable and neuroanatomy defective. *Abl* is discussed in 206 references (excluding sequence accessions), dated between 1981 and 2005. These include at least 30 studies of mutant phenotypes , 8 studies of wild-type function and 10 molecular studies . Among findings on *Abl* mutants, *Abl* mutants show phenotypes in somatic muscles and eye imaginal disks. Among findings on *Abl* function, *Abl* gene product may play a role in establishing and maintaining cell-cell interactions.

result, most sentences in *MP* and *GI* categories will be judged as irrelevant. Thus, the low precision in some occasions may simply be because there is little research on this topic. In general, the lack of information on some

Table 5. Text summary of gene *Camo\Sod* by our system

GP	Superoxide production by Drosophila mitochondria was measured fluorometrically as hydrogen peroxide, using its dependence on substrates, inhibitors, and added superoxide dismutase to determine sites of production and their topology.
EL	The aim of this study was to ascertain the status of CuZn superoxide dismutase (CuZn-SOD) expression in the central nervous system of Drosophila melanogaster.
SI	Comparison of the Drosophila Cu,Zn SOD amino acid sequences with the Cu,Zn SOD of Bos taurus and Xenopus laevis (whose three-dimensional structure has been elucidated) reveals conservation of all the protein's functionally important amino acids and no substitutions that dramatically change the charge or the polarity of the amino acids.
MP	The gene for cytoplasmic superoxide dismutase (cSOD) maps within this interval, as does low xanthine dehydrogenase (lxd).–Recessive lethal mutations were generated within the region by ethyl methanesulfonate mutagenesis and by hybrid dysgenesis.
GI	Drosophila orthologues of the mammalian Cu chaperones, ATOX1 (a human orthologue of yeast ATX1), CCS (copper chaperone for superoxide dismutase), COX17 (a human orthologue of yeast COX17), and SCO1 and SCO2, did not significantly respond transcriptionally to increased Cu levels, whereas MtnA, MtnB and MtnD (Drosophila orthologues of human metallothioneins) were up-regulated by Cu in a time- and dose-dependent manner.
WFPI	The 2.5 kb clone consists of a wild-type 1.84 kb EcoRI fragment containing the Cu,Zn SOD gene previously isolated in our laboratory, with an insertion of 0.68 kb derived (by an internal deletion) from an autonomous, 2.9 kb P element.

Table 6. Text summary of gene *Camo\Sod* from FlyBase

Superoxide dismutase, abbreviated as ***Camo\Sod***, is reported here. It has been sequenced . There is one recorded allele, which is wild-type. *Camo\Sod* is discussed in 4 references (excluding sequence accessions), dated between 1992 and 2001.

aspects of a query gene is not a major problem for our system in the sense that, if information about one aspect is missing, a biologist could infer that this aspect may have not been well studied or is not biologically interesting.

5. Conclusion and Future Work

In this paper, we proposed a novel problem in biomedical text mining: automatic generation of structured gene summaries. We developed a system which employed information retrieval and information extraction techniques to automatically summarize information about genes from PubMed abstracts. The system was tested on 10 randomly selected genes, and evaluated by domain experts. The promising results with an average precision above 50% indicate that the system is very effective in summarizing biomedical literature.

We realized that one obvious limitation of our approach was its dependence on the high-quality data in FlyBase. To address this issue, we will in the future incorporate more training data from databases of other model organisms and resources such as GeneRIF in Entrez Gene. We believe the mixture of data from different resources will reduce the domain bias and help build a general tool for gene summarization.

We employed many heuristic methods in our system, primarily because it is unclear at the beginning which computational strategy would be most suitable for our problem. A major future work is to explore more generic methods including probabilistic models for sentence selection. Our long-term goal is to extend our system so that it can be used by all biomedical researchers. Even though we used some fly-specific resources and tested mainly on fly genes, the general framework we proposed is independent of the actual biological domains. We will next be testing the methods on bee genes using the same training set on fly genes but extracting sentences from bee literature, to test applicability across insects. Eventually, we hope to produce automatic summarization of all genes in all organisms, using the entire biomedical literature for extraction and the entire set of model genetic databases for training.

References

1. I. Iliopoulos, A. Enright, C. Ouzounis, (2001) Textquest: document clustering of medline abstracts for concept discovery in molecular biology. *PSB*, 384-395.
2. L. Hirschman, J. C. Park, J. Tsujii, L. Wong, C. H. Wu, (2002) Accomplishments and challenges in literature data mining for biology. Bioinformatics **18(12)**:1553-1561.
3. H. Shatkay, R. Feldman, (2003) Mining the Biomedical Literature in the Genomic Era: An Overview. *JCB.*, **10(6)**:821-856.
4. D. Marcu, (2003) Automatic Abstracting. *Encyclopedia of Library and Information Science*, 245-256.
5. D. R. Radev, H. Jing, M. Sty, D. Tam, (2004) Centroid-based summarization of multiple documents. Inf. Process. Manage. 40(6):919-938.
6. L. Hirschman, M. Colosimo, A. Morgan, A. Yeh, (2005) Overview of BioCreAtIvE Task 1B: Normailized Gene Lists. *BMC Bioinformatics* 2005, 6(Suppl):S11.
7. R. A. Drysdale, M. A. Crosby and The FlyBase Consortium, (2005) FlyBase: genes and gene models. *Nucleic Acids Res.* **33**: 390-395.