Validating Candidate Gene-Mutation Relations in MEDLINE Abstracts via Crowdsourcing

John D. Burger[1], Emily Doughty[2], Sam Bayer[1], David Tresner-Kirsch[1], Ben Wellner[1], John Aberdeen[1], Kyungjoon Lee[3], Maricel G. Kann[2], Lynette Hirschman[1]

[1]The MITRE Corporation, Bedford, MA, USA
{john, sam, davidtk, wellner, aberdeen, lynette}@mitre.org
[2]University of Maryland, Baltimore County, Baltimore MD, USA
{doughty2, mkann}@umbc.edu
[3]Harvard Medical School, Boston MA, USA
joon_lee@hms.harvard.edu

Abstract. We describe an experiment to elicit judgments on the validity of gene-mutation relations in MEDLINE abstracts via crowdsourcing. The biomedical literature contains rich information on such relations, but the correct pairings are difficult to extract automatically because a single abstract may mention multiple genes and mutations. We ran an experiment presenting candidate gene-mutation relations as Amazon Mechanical Turk *HITs* (human intelligence tasks). We extracted candidate mutations from a corpus of 250 MEDLINE abstracts using EMU combined with curated gene lists from NCBI. The resulting document-level annotations were projected into the abstract text to highlight mentions of genes and mutations for review. Reviewers returned results within 36 hours. Initial weighted results evaluated against a gold standard of expert curated gene-mutation relations achieved 85% accuracy, with the best reviewer achieving 91% accuracy. We expect performance to increase with further experimentation, providing a scalable approach for rapid manual curation of important biological relations.

Keywords: Crowdsourcing, text mining, metadata curation, annotation

1 Introduction

The decrease in genome sequencing costs has led to an explosion of information on mutations in the human genome and their relation to disease. There are now a number of curated databases devoted to the capture of information on mutations in specific populations with associated phenotypes, including, e.g., OMIM [1], dbSNP [2], HGVbase [3], HGMD [4] and locus specific databases (LSDBs) [5]. Such databases are critical for downstream data integration, making it possible, for example, to interpret patient genetic data (SNPs in genes) against variants known to be associated with disease. These resources are also critical for genome-wide association studies (GWAS) [6], phenotype-wide association studies (PheWAS) [7] and pharmacogenomics [8]. However, much of the information about these gene-mutation-disease associations is currently buried in the biomedical literature. There is increased demand to identify these results in a timely fashion and to make them available in a computable form to support personalized medicine and translational medicine applications.

In this paper, we focus on the curation challenges for a subportion of this problem: identifying relationships between genes and mutations from the biomedical literature. We investigate the efficacy of *crowdsourcing* to validate gene-mutation relations in MEDLINE abstracts. Crowdsourcing is the process of leveraging a large potential pool of non-experts for problem solving via the Web. Here we employ crowdsourcing to verify whether candidate gene-mutation associations in MEDLINE abstracts represent valid associations, using automated techniques to identify candidate genes and mutations in abstracts for presentation to reviewers.

Identifying relevant associations automatically in the literature requires two steps: identifying the elements (genes and mutations), and determining which ones are related. For MEDLINE abstracts, one of these sets of elements (genes) is routinely annotated by indexers at the National Library of Medicine, but mutation terms must be extracted directly from the abstract text, generally via regular expressions and/or natural language processing methods.

Mutations are challenging to identify in the literature because of the many types of mutations (SNPs, indels, rearrangements) and the variability in how the mutations are described in the literature. For example, SNPs (single nucleotide polymorphisms), the simplest class of mutation, can be described as a mutation in the DNA (in terms of nucleotides) or as a mutation in the protein (in terms of amino acids). Examples of SNPs found in the literature include condensed forms, such as “313A>G”, “E161K”, “Pro582Ser” or “AGG to AGT”, as well as whole phrases, such as “substitution of Met-69 by Ala or Gly in TEM-1 beta-lactamase”. In spite of these complexities, locating a subset of these mutations (especially SNPs) in the literature has been shown to be a tractable task for text-mining tools based on pattern matching, as shown by MutationFinder [9], Extractor of MUtations (EMU) [10], MutationTagger [11] and elsewhere [12-14].

The focus of the experiment described here is to establish the correct association between a mutation and the associated gene, otherwise known as *mutation grounding* [11]. When multiple mutations and genes are found in the same text, the association of mutations to genes is challenging, leading to false positives. Using EMU to extract mutations from abstracts, Doughty et al. [10] reported a high precision on detection of the mutation patterns (99 and 96% for prostate and breast cancer mutations, respectively) but a significant decrease in precision, of up to 20%, when attempting to automatically extract the correct gene-mutation pairs. Mutation grounding can be improved by filtering based on match of wildtype sequence to a reference sequence, given gene and positional information, as described in [10, 11]; however, this results in a decrease in recall. For these reasons, the mutation grounding task seemed like an interesting candidate for crowdsourcing.

2 Data

A gold standard gene-mutation-disease data set was created for three diseases (breast cancer, prostate cancer, and autism spectrum disorder) as follows. Mutation-related MEDLINE abstracts were downloaded from the PubMed search engine using the MeSH terms “Mutation” (breast and prostate cancer) or “Mutation AND Polymorphism/genetic” (autism spectrum disorder). Abstracts related to breast (5967 citations) and prostate cancer (1721 citations) were identified by MetaMap [15] as described in [10]. An additional set of abstracts related to autism spectrum disorder were identified using the MeSH term “Child Development Disorders, Pervasive.” The EMU tool was used to identify mutations in the three disease abstract corpora. Abstracts for which EMU identified at least one mutation were selected for expert curation, resulting in a corpus of 810 MEDLINE abstracts, with 1573 mutations; almost 50% of the abstracts had two or more curated gene-mutation relationships.

For the initial crowdsourcing experiment, we focused only on the gene-mutation relations in the gold standard. Because of limitations of time and budget, we used a randomly selected 250-abstract subset of the full gold standard. Of these, six contained no mutations, 99 contained a single mutation, and the remaining 145 contained two or more mutations. The subset contained 568 gold-standard mutations altogether.

3 Methods

3.1 Overall Experimental Design

We designed the gene-mutation relation validation as a two-stage process. The first stage automatically identified all candidate mutations and genes in each abstract and then projected these candidates back into the text to highlight the specific mentions. The second stage utilized Amazon Mechanical Turk, an online crowdsourcing platform, to elicit judgments from multiple reviewers on correct gene-mutation pairings.

In stage 1, we used a modified version of EMU to extract and normalize mutations. For genes, we took the union of the NLM-curated genes associated with each abstract in PubMed and any additional genes extracted by EMU. We then generated all possible gene-mutation pairs for each abstract and constructed a separate item for each pair in each abstract for presentation to the reviewers. Each item is a yes/no question asking whether the abstract discusses the candidate relation between the highlighted gene and mutation. In stage 2, the items were distributed via Amazon Mechanical Turk, and the results were aggregated to provide judgments on the validity of each presented gene-mutation pair. To evaluate, we compared the aggregated reviewer judgments on gene-mutation pairs to the gold standard data. We also compared the output of the first stage pre-processing to the gold standard data, to determine loss of recall in preparing the data, and amount of over-generation.

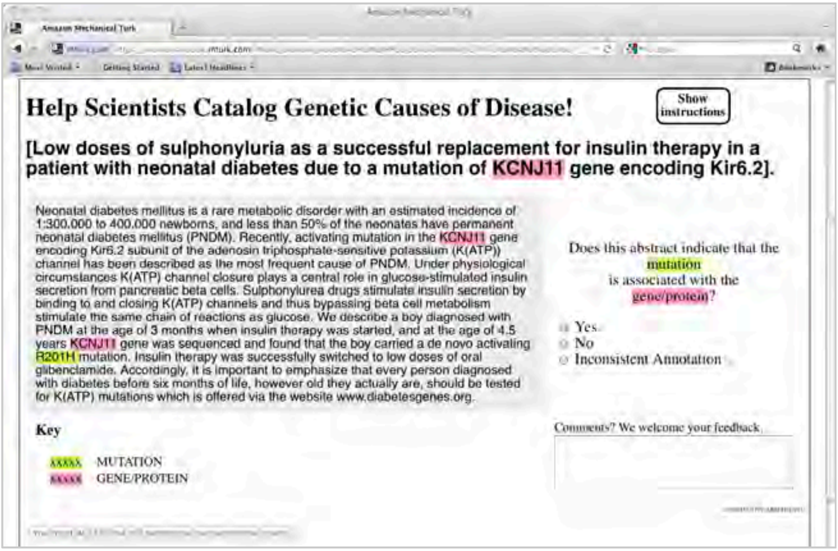
3.2 Extraction of Candidate Genes, Mutations and Their Mentions

EMU [10] extracts point mutations from text and identifies relevant genes found in the input text. Gene identification is done using string matching against a customized list of gene names derived from the Human Genome Organization (HUGO) and NCBI Gene databases. To ensure maximal coverage of potentially relevant gene names, we augmented the genes identified by EMU with any additional genes curated for each article by NCBI. We then found all occurrences of these genes in the input text using exact string matching. In the case of NCBI genes, we searched for occurrences of the gene symbol, the gene name and any synonyms for the gene that appeared in the database. For mutations, EMU detects SNPs through a two-step filter. The first step collects likely mutation spans using a list of positive regular expressions; the second rejects candidates based on a stop-list of negative regular expressions. For both genes and mutations, we identified the text spans in the input text associated with each gene mention. To combine and visualize these tagged text spans, we relied on tools for human linguistic annotation found in the MITRE Identification Scrubber Toolkit [16]. These tools allowed us to create rich documents with *standoff annotations* that identify the type and location of the mention. Standoff annotations are not embedded in the text, and are thus amenable to manipulation and processing. The toolkit also provided a browser-based presentation library for highlighting the location of these mentions.

3.3 Amazon Mechanical Turk for Gene-Mutation Association Identification

Amazon Mechanical Turk (MTurk) is a web-based labor market for *Human Intelligence Tasks* (*HITs*). HITs are typically minimal problems that cannot easily be automated, such as categorizing a photograph or a text snippet. Most HITs take a few seconds for a worker to perform, and pay a few cents. In 2011, Amazon indicated that there were over 500,000 workers (Turkers) registered, although as with all online services, many more people sign up than are active at any time. A number of researchers have recently experimented with the use of MTurk to create and annotate human language data [17]. In particular, MTurk has been used to annotate medical named entities in clinical trials descriptions [18].

To prepare candidate gene-mutation pairs for presentation to the Turkers, we first grouped together multiple mentions of the same gene, and also multiple mentions of each mutation. For genes, we used the gene ID, or if unavailable, the gene name. Mutations were canonicalized according to EMU by a triple of position, wild type and mutant nucleotide or amino acid. We took the cross product of all genes and all mutations found in an abstract, which resulted in 1299 pairs. Each of these gene-mutation candidates was presented as a separate HIT on MTurk. For each HIT, the Turker was asked a yes-or-no question to determine whether the given mutation and gene are in fact related. (Turkers were also allowed to choose *inconsistent annotation* to indicate a problem in the projection—this option was rarely used, but was counted as a *no* answer.) The mentions of the gene and mutation in the relevant pair were highlighted in the abstract's text using the projection method described above. Figure 1 shows a



Help Scientists Catalog Genetic Causes of Disease!

[Low doses of sulphonyluria as a successful replacement for insulin therapy in a patient with neonatal diabetes due to a mutation of KCNJ11 gene encoding Kir6.2].

Neonatal diabetes mellitus is a rare metabolic disorder with an estimated incidence of 1:300,000 to 400,000 newborns, and less than 50% of the neonates have permanent neonatal diabetes mellitus (PNDM). Recently, activating mutation in the KCNJ11 gene encoding Kir6.2 subunit of the adenosin triphosphate-sensitive potassium (K(ATP)) channel has been described as the most frequent cause of PNDM. Under physiological circumstances K(ATP) channel closure plays a central role in glucose-stimulated insulin secretion from pancreatic beta cells. Sulphonylurea drugs stimulate insulin secretion by binding to and closing K(ATP) channels and thus bypassing beta cell metabolism stimulate the same chain of reactions as glucose. We describe a boy diagnosed with PNDM at the age of 3 months when insulin therapy was started, and at the age of 4.5 years KCNJ11 gene was sequenced and found that the boy carried a de novo activating R201H mutation. Insulin therapy was successfully switched to low doses of oral glibenclamide. Accordingly, it is important to emphasize that every person diagnosed with diabetes before six months of life, however old they actually are, should be tested for K(ATP) mutations which is offered via the website www.diabetesgenes.org.

Does this abstract indicate that the mutation is associated with the gene/protein?

○ Yes
○ No
○ Inconsistent Annotation

Key

🟩 MUTATION
🟥 GENE/PROTEIN

Comments? We welcome your feedback.

Figure 1: HIT design for the gene-mutation task

screen capture of the HIT design, in which a single gene-mutation pair candidate is presented.

Because Turkers are typically not pre-qualified for any particular task, there are several strategies that requesters use to screen Turkers for tasks requiring specific expertise or skills. The requester can require that Turkers be from a particular region, or have a certain minimum approval rating on their previous HITs. In addition, the requester can leverage redundancy by asking for multiple Turkers to perform each HIT. The responses can then be aggregated in some way, perhaps by simply choosing the majority response. Another strategy is to insert *control items* into the HITs. These are items for which the correct answer is already known. Each Turker has a persistent ID that accompanies their responses, and by aggregating control item responses, each Turker's efficacy can be measured.

We used all of these strategies to control for Turker variability. First, we required the Turkers for these HITs to be from the United States, in order to minimize second language effects. We also limited participation to Turkers with 95% approval ratings, and we requested that five separate Turkers annotate each HIT. Finally, we included a number of control items. These were 44 gene-mutation pairs from a handful of separate abstracts developed in initial annotation experiments. The control items were identified as relatively easy to annotate—that is, even non-experts should agree as to whether a candidate pair was positive or negative. We created multiple HITs for each control item such that approximately 25% of the HITs each Turker encountered were these known items. By injecting these control items into the HITs, we are able to estimate the efficacy of the Turkers as described above.

4 Results

We measured how well the pre-processing captured the relevant mutations and genes by comparing them independently against the sets of mutations and genes that appeared in the gold-standard curated gene-mutation associations. In the 250 abstracts in our dataset, there were 568 unique mutations at the abstract level in the gold standard; the pre-processing via EMU correctly extracted 484 (recall of 85.2%). There were 279 unique genes at the abstract level; the gene pre-processing using EMU and NCBI gene lists extracted 260 of these genes (recall of 93.2%).

Overall, there were 1299 candidate gene-mutation pairs presented to the Turkers, and 568 unique mutation-gene pairs in the gold standard, giving a distribution of 44:56 positive to negative candidates. Of the 250 abstracts, there were 181 for which *all* of the mutations and genes that appeared in the curated gene-mutation pairs were identified by our automated methods. These abstracts are of particular interest for our experiments since all of the curated gene-mutation relations for these documents could, in theory, be identified through crowdsourcing. We refer to these below as *perfect documents*.

1733 HITs were posted to MTurk, including 1299 candidate gene-mutation pairs generated from the 250 abstracts, with the rest being control items as described above. We requested that five Turkers work on each HIT, for a total of 8665 judgments. These were completed within 36 hours of posting them. Altogether, 23 Turkers responded, although seven only performed one HIT. Eight Turkers performed ten or more HITs, and six performed 100 or more. We paid 7¢ per judgment, and with Amazon's overhead fees the session cost approximately $670 (U.S.).

In Table 1, we show Turker accuracy on the 1299 test items in the All Docs column. The average response accuracy is 76.5%, which is higher than a random baseline. The average accuracy across Turkers is lower, due to a number of Turkers who only did a few HITs and performed poorly. As we look at successively higher thresholds for number of responses, we see increases in accuracy, with the best Turker performing at an accuracy of 90.5%. We also show Turker performance on the candidates drawn from the 181 perfect documents described above (Table 1, Perfect Docs). Comparing the results, the Turkers show some improvement (around 2 percentage points) on these documents.

We can aggregate the annotations from multiple Turkers to produce a consensus judgment for each HIT, in order to compensate for poor performers. For example, a majority vote approach counts each Turker judgment as a vote for the correct label for

	All Docs	Perfect Docs
Average response	76.5%	79.3
Average Turker	62.0	68.7
Average 10+ Turker	70.7	73.2
Average 100+ Turker	76.0	78.6
Best Turker	90.5	92.4

Table 1: Accuracy for overall Turker responses and various subsets

that HIT (candidate gene-mutation relation). The accuracy for a simple majority approach over all documents (83.8%) is substantially better than the average response performance (76.5%). Another approach is to use each Turker's control item accuracy to weight their vote, capitalizing on the intuition that the opinion of Turkers who score well on the control items should count for more. However, this performs no better than the simple majority approach (83.7%). More principled approaches include combining labels from different Turkers using probabilistic classifiers. We can treat each Turker response as an observed feature on the candidate item, with the control items viewed as training data. We applied two such frameworks to the data: a Naïve Bayes classifier achieved 84.5% accuracy, while logistic regression performed at 83.2% accuracy.

5 Discussion

Our initial results are promising. We had no difficulty recruiting qualified Turkers and—of particular note—they returned results remarkably quickly (36 hours from release of the HITs). The best Turker achieved a very respectable accuracy of 90.5%, annotating 1144 of 1264 HITs correctly.

Based on a preliminary error analysis, we believe that the reported results underestimate Turker performance. We reviewed 200 cases (ranked by confidence) where Turker aggregate judgments differed from the gold standard. For 50 HITs, there was a mismatch between the gold standard and EMU (v1.0.16) in treatment of SNPs in non-coding regions of genes. In addition, in some 20 HITs, there was a problem with normalization in the gold standard that caused a mismatch. This suggests that overall Turker accuracy may be significantly higher once such discrepancies are resolved, which we plan to do in our follow up experiments.

Feedback from Turkers leads us to believe that the interface provided an effective way to present annotation decisions to reviewers. Because the Turker/reviewer is presented with a binary decision with the relevant context made salient (by highlighting the gene and mutation candidates in the text), these decisions can be made quite quickly. This particular task turned out to be well-suited to the use of Amazon Mechanical Turk, but we recognize that there may be significant limitations to the crowdsourcing approach for more complex tasks, such as gene-mutation-disease relations, where deep subject matter expertise is needed. We also note that for this experiment, per-abstract costs were significantly higher than curation done by smart undergraduates: $2.50 (U.S.) per abstract done by five Turkers with aggregate accuracy of 85% vs. $0.50 per abstract for undergraduates at an average accuracy of 92%. However, we expect that these costs can be reduced by refinements in distribution of HITs to Turkers and more sophisticated ways of vetting Turkers and aggregating their results.

Finally, we can make a rough estimate of the importance of the biomedical literature as a source of novel findings for human genetic variants. Of the 1770 explicit gene-mutation relations found from the 810 expert-curated abstracts, two thirds (1163) did not appear in OMIM, Swiss-Prot [19] or dbSNP, suggesting that the literature is a rich source of novel gene-mutation relations, even given the 85% recall of the tool chain.

6 Conclusions

Scalable timely capture of gene-mutation-disease information is of critical importance for the rapidly growing field of personalized medicine. The biomedical literature remains a rich source of such information, and validation of relations extracted from the literature is an important step in this process. We have presented a promising initial experiment using crowdsourcing to validate gene-mutation relations assembled from automatically extracted genes and mutations. We were able to recruit high-performing Turkers; they returned results within a day and a half and provided positive feedback on the task and the interface. This suggests that crowdsourced judgments on the validity of candidate biological relations may provide a scalable rapid turn-around approach to obtaining such information.

Acknowledgements. MGK and ED were supported by National Institutes of Health (NIH) Grant 1K22CA143148 to MGK. KL was supported by NIH grant P50MH094267.

References

1. Amberger, J., C.A. Bocchini, A.F. Scott and A. Hamosh (2009) *McKusick's Online Mendelian Inheritance in Man (OMIM).* Nucleic Acids Res. **37**(Database issue): p. D793-6.
2. Sherry, S.T., M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, and K. Sirotkin (2001) *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res. **29**(1): p. 308-11.
3. Thorisson, G.A., O. Lancaster, R.C. Free, R.K. Hastings, P. Sarmah, D. Dash, S.K. Brahmachari, and A.J. Brookes (2009) *HGVbaseG2P: a central genetic association database.* Nucleic Acids Res. **37**(Database issue): p. D797-802.
4. Stenson, P.D., E.V. Ball, K. Howells, A.D. Phillips, M. Mort, and D.N. Cooper (2009) *The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics.* Human genomics. **4**(2): p. 69-72.
5. Samuels, M.E. and G.A. Rouleau (2011) *The case for locus-specific databases.* Nat Rev Genet. **12**(6): p. 378-379.
6. Klein, R.J., C. Zeiss, E.Y. Chew, J.Y. Tsai, R.S. Sackler, C. Haynes, A.K. Henning, J.P. SanGiovanni, S.M. Mane, S.T. Mayne, M.B. Bracken, F.L. Ferris, J. Ott, C. Barnstable, and J. Hoh (2005) *Complement factor H polymorphism in age-related macular degeneration.* Science. **308**(5720): p. 385-9.
7. Denny, J.C., M.D. Ritchie, M.A. Basford, J.M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D.R. Masys, D.M. Roden, and D.C. Crawford (2010)

PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. **26**(9): p. 1205-10.

8. Tatonetti, N.P., J.T. Dudley, H. Sagreiya, A.J. Butte and R.B. Altman (2010) *An integrative method for scoring candidate genes from association studies: application to warfarin dosing.* BMC Bioinformatics. **11 Suppl 9**: p. S9.
9. Caporaso, J.G., W.A. Baumgartner, Jr., D.A. Randolph, K.B. Cohen and L. Hunter (2007) *MutationFinder: a high-performance system for extracting point mutation mentions from text.* Bioinformatics. **23**(14): p. 1862-5.
10. Doughty, E., A. Kertesz-Farkas, O. Bodenreider, G. Thompson, A. Adadey, T. Peterson, and M.G. Kann (2011) *Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature.* Bioinformatics. **27**(3): p. 408-415.
11. Winnenburg, R., C. Plake and M. Schroeder (2009) *Improved mutation tagging with gene identifiers applied to membrane protein stability prediction.* BMC Bioinformatics. **10**(Suppl 8): p. S3.
12. Rebholz-Schuhmann, D., S. Marcel, S. Albert, R. Tolle, G. Casari, and H. Kirsch (2004) *Automatic extraction of mutations from Medline and cross-validation with OMIM.* Nucleic Acids Res. **32**(1): p. 135-42.
13. Horn, F., A.L. Lau and F.E. Cohen (2004) *Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors.* Bioinformatics. **20**(4): p. 557-68.
14. Erdogmus, M. and O.U. Sezerman (2007) *Application of automatic mutation-gene pair extraction to diseases.* J Bioinform Comput Biol. **5**(6): p. 1261-75.
15. Aronson, A.R. (2001) *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.* Proc AMIA Symp: p. 17-21.
16. Aberdeen, J., S. Bayer, R. Yeniterzi, B. Wellner, C. Clark, D. Hanauer, B. Malin, and L. Hirschman (2010) *The MITRE Identification Scrubber Toolkit: design, training, and assessment.* International journal of medical informatics. **79**(12): p. 849-59.
17. Callison-Burch, C. and M. Dredze (2010). *Creating speech and language data with Amazon's Mechanical Turk.* in *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Los Angeles, California: Association for Computational Linguistics.
18. Yetisgen-Yildiz, M., I. Solti, F. Xia and S. Halgrim (2010) *Preliminary Experiments with Amazon's Mechanical Turk for Annotating Medical Named Entities*, in *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Assn for Comp. Ling.: Los Angeles. p. 180-183.
19. Boeckmann, B., A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider (2003) *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res. **31**(1): p. 365-70.