Learning string similarity measures for gene/protein name dictionary look-up using logistic regression

Yoshimasa Tsuruoka¹,*, John McNaught¹,², Jun’ichi Tsujii¹,²,³ and Sophia Ananiadou¹,²

¹School of Computer Science, The University of Manchester, Manchester, ²National Centre for Text Mining (NaCTeM), Manchester, UK and ³Department of Computer Science, The University of Tokyo, Japan

Received on June 3, 2007; revised on July 27, 2007; accepted on July 28, 2007
Advance Access publication August 12, 2007
Associate Editor: Jonathan Wren

1 INTRODUCTION

Looking up a gene/protein dictionary is a common task for both computer systems and researchers in biomedical research. Many of the information extraction systems developed for biomedical documents provide a mapping between gene/protein names found in text and their corresponding identifiers (IDs) in the database. In a biological database, due to seemingly small differences of names, many string matching methods are employed. A dictionary look-up system using the similarity measure obtained through experimental use of dictionary look-up tasks. Another approach, which may be employed in conjunction with the normalization approach, is to use soft string matching methods. Soft matching gives similarity scores between strings, which allows us to associate term forms even when they are not identical. Moreover, soft matching can provide the user with multiple candidates that are ranked according to their similarity scores.

The effectiveness of soft matching almost exclusively depends on the design of the similarity measure that quantifies the degree of similarity between two given strings. One could use a similarity measure designed manually (krauthammer et al., 2000; Tsuruoka and Tsujii, 2004), but recent studies have shown that automatically tuned measures often give better results. For the task of associating gene/protein names, Yeganova et al. (2004) used a hidden Markov model which optimizes its parameters by using synonymous pairs of strings in the dictionary. Cohan and Minkov (2006) used a soft-matching technique called SoftTFIDF, which enables us to focus on salient words when comparing the strings.

This article explores the use of logistic regression to learn a good string similarity. Unlike the aforementioned method based on hidden Markov models, we use not only synonymous pairs of strings but also non-synonymous pairs when optimizing the similarity measure. Moreover, the model allows us to make use of diverse types of information as features that characterize each string pair.

We compare the performance of several similarity measures against the one we derive through logistic regression in two sets of dictionary lookup experiments. In the first set of experiments, we use five species-specific dictionaries, and evaluate the performance of similarity measures using the entries which are held out from each dictionary. In the second set of experiments, we use gene/protein names that actually appear in MEDLINE abstracts for evaluation.

ABSTRACT

Motivation: One of the bottlenecks of biomedical data integration is variation of terms. Exact string matching often fails to associate a name with its biological concept, i.e. ID or accession number in the database, due to seemingly small differences of names. Soft string matching potentially enables us to find the relevant ID by considering the similarity between the names. However, the accuracy of soft matching highly depends on the similarity measure employed.

Results: We used logistic regression for learning a string similarity measure from a dictionary. Experiments using several large-scale gene/protein name dictionaries showed that the logistic regression-based similarity measure outperforms existing similarity measures in dictionary look-up tasks.

Availability: A dictionary look-up system using the similarity measures described in this article is available at http://text0.mib.man.ac.uk/software/mldic/

Contact: yoshimasa.tsuruoka@manchester.ac.uk

*To whom correspondence should be addressed.

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
This article is organized as follows. Section 2 summarizes previous work on string similarity measures. In Section 3, we describe the dictionaries and data sets used in the experiments. Section 4 presents a similarity measure based on logistic regression and features used for representing a sample. Section 5 describes experimental results using dictionaries created from BioThesaurus (Liu et al., 2006) and the data provided by the BioCreAtIvE II Gene Normalization task (Morgan and Hirschman, 2007).

2 RELATED WORK

Simple similarity measures for soft string matching include character n-gram similarity, the Levenshtein distance (Levenshtein, 1965) and the Jaro–Winkler measure (Winkler, 1999), in which the same penalty value is used regardless of the characters to be matched (or ignored). In general, these simple measures do not work very well for gene/protein names on their own, but they can be useful in situations where the computational cost needs to be minimized.

One can employ a more sophisticated measure by defining different scores for different characters and matching operations. It is straightforward to use different scores in edit distance-like similarity measures. Karuthammer et al., (2000) used the well-known BLAST algorithm to identify gene and protein names in text. They converted each character in the strings into an amino-acid sequence so that the BLAST algorithm can find matched strings in a sentence using the similarity measure originally developed for gene sequence alignment. Tsuruoka and Tsuji, (2004) used an edit distance measure for protein named entity recognition. They manually tuned the cost function to make it less sensitive to capitalization and hyphenation, and reported an improvement of recall.

There are algorithms that enable us to “learn” string similarity from actual examples of string pairs. Ristad and Yianilos (1998) proposed a generative model for edit distance, and presented an algorithm for tuning the cost of edit operations using synonymous pairs of strings. Smith et al. (2003) proposed to use an HMM-based probabilistic model for sequence alignment, and described a training algorithm based on the forward-backward algorithm with which one can estimate the parameters of the HMM using pairs of relevant sequences as the training data. Yeganova et al. (2004) applied this model to the identification of related gene/protein names using manually curated training data, and reported their advantages over the aforementioned BLAST-based method. Wellner et al. (2005) proposed to train conditional random fields on the optimal operation sequences given by the Levenshtein distance.

The learnable string similarity approaches described above, however, use only synonymous pairs of strings for tuning the parameters. A relatively new line of research is to use non-synonymous pairs as well as synonymous pairs by employing discriminative learning models. Cohen and Richman (2002) proposed to use a maximum entropy-based binary classifier to combine multiple similarity metrics, and applied their method to the task of integrating database entities. Bilenko and Mooney (2003) used a support vector machine for a similar task. Bilenko et al. (2005) proposed online learning of similarity functions using a voted-perceptron algorithm. McCaCallum et al. (2005) presented a string edit distance function based on conditional random fields, which allows us to use a variety of features of strings and edit operations.

These discriminative learning-based approaches are attractive because they (a) enable us to explicitly consider the ‘dissimilarity’ of strings, (b) allow us to incorporate a variety of features for characterizing a string pair. For instance, the protein names ‘GATA binding protein 2’ and ‘GATA binding protein 5’ are very similar on the character level, but one might want to focus on the difference in the numbers (‘2’ and ‘5’) when quantifying the similarity between them. This type of information can be easily incorporated as a feature in these discriminative learning models.

3 DATA FOR TRAINING AND EVALUATION

This work is largely motivated by the recent development of large-scale gene/protein name dictionaries, including GENA (Koike et al., 2003), ProMiner (Hanisch et al., 2005), and BioThesaurus. These dictionaries are typically constructed by extracting names and descriptions from general biological databases (e.g. HUGO, OMIM, Swiss-Prot, Locuslink) and species-specific databases (e.g. MGI, FlyBase, MGD, SGD).

What makes these dictionaries particularly appealing is that they contain variants of names as well as canonical names. Table 1 shows an example of a gene/protein dictionary. Actual dictionaries typically contain other types of information such as DNA sequences and literature references, but here we focus only on the names and their IDs.

We can learn a variety of information from such dictionary entries. For example, Table 1 tells us that we should match ‘AP3BI’ and ‘AP-3 complex subunit beta-1’ because these names share the same ID. At the same time, the dictionary tells us that we should not match ‘AP-3 complex subunit beta-1’ with ‘AP-3 complex subunit mu-2’ because they belong to different IDs. In other words, these entries suggest that we treat ‘AP3BI’ and ‘AP-3 complex subunit beta-1’ as being similar, and ‘AP-3 complex subunit beta-1’ and ‘AP-3 complex subunit mu-2’ as being dissimilar.

BioThesaurus (Liu et al., 2006) is a collection of more than two million gene/protein names from many different species, and as it includes variants of different types it is a useful resource to enable us to learn string similarity.

For the experiments in this article, we create species-specific dictionaries from BioThesaurus data1 for five species (Human, Mouse, E.coli, Yeast and Drosophila). Each entry in BioThesaurus is associated with a UniProt ID. We consulted the UniProtKB/Swiss-Prot database2 for selecting a species-specific subset from BioThesaurus. We then removed nonsensical terms, e.g. accession numbers for other databases, using simple regular expressions. Each resulting dictionary is split into two sets. One is used for training, and the other is used for evaluation.

For the second set of experiments, we use a human gene/protein name dictionary, and names that actually appear in MEDLINE abstracts. We extracted these data from the data

---

1^bio Thesaurus.dist_2.0.gz^ available at ftp.pir.georgetown.edu.
2^Available at http://www.ebi.uniprot.org/database/download.shtml
set originally developed for the gene normalization task at BioCreAtIvE II (2006). The original data set contains the text of MEDLINE abstracts as well, but we do not use them since the recognition of gene/protein names is not our focus in this article.

Table 2 shows statistics of the dictionaries used in the experiments. The dictionaries contain many variants. Each ID has 5–14 synonymous names on average.

4 LEARNING STRING SIMILARITY MEASURES USING LOGISTIC REGRESSION

As discussed in the previous sections, a dictionary provides information about what kind of string pairs should be regarded as synonymous (or non-synonymous). In order to build a similarity measure based on the dictionary, we need to generalize this information so that we can make this judgement (synonymous or not) for an arbitrary pair of strings.

One way of achieving this generalization is to use a machine-learning approach. In particular, we can use supervised classification models. The task is formalized as a binary classification problem, where the input is a pair of strings and the output is a prediction of whether the strings are synonymous or not. In general, a machine learning-based classifier can output a confidence value for each prediction, so we should not expect a machine-learning algorithm, which solely relies on surface string similarity, to associate the names.

To let the logistic regression model learn from only meaningful samples, we introduce a filtering process. We create a training sample from a string pair, whether it is synonymous or not, only when at least one of the following conditions is satisfied:

- The two strings have a high value (>0.5) of character bigram similarity, which is computed as follows:

\[
\text{similarity} = \frac{2g_1 \cap g_2}{|g_1| + |g_2|},
\]

where \(g_1\) and \(g_2\) are the bigrams in the strings.

- All the characters in the shorter string are included in the longer string in the same order.

In the evaluation stage, the measure simply gives a similarity value of 0 to the pairs which do not pass this filtering.

This filtering process also has the merit of reducing the amount of training samples, but the training cost was still very high. The number of training samples for non-synonymous pairs is much higher than that for synonymous pairs, and we found, in preliminary experiments, that reducing the samples for non-synonymous pairs does not have a large impact on the overall performance. We therefore discarded three quarters of the non-synonymous samples by random sampling.3

Table 1. Part of a gene/protein name dictionary

<table>
<thead>
<tr>
<th>ID</th>
<th>Gene/protein name</th>
</tr>
</thead>
<tbody>
<tr>
<td>O00203</td>
<td>AP3B1</td>
</tr>
<tr>
<td>O00203</td>
<td>AP-3 complex subunit beta-1</td>
</tr>
<tr>
<td>O00203</td>
<td>AP-3 complex subunit beta-1</td>
</tr>
<tr>
<td>O00203</td>
<td>Adapter-related protein complex 3</td>
</tr>
<tr>
<td>O00203</td>
<td>Adapter protein complex AP-3 beta-1</td>
</tr>
<tr>
<td>O00203</td>
<td>Beta-3A-adaptin subunit of the AP-3 complex</td>
</tr>
<tr>
<td>O00203</td>
<td>HPS</td>
</tr>
<tr>
<td>O00203</td>
<td>HSP2</td>
</tr>
<tr>
<td>O00203</td>
<td>HPS2 GENE</td>
</tr>
<tr>
<td>O00203</td>
<td>HERMANSKY-PUDLAK SYNDROME</td>
</tr>
<tr>
<td>O00203</td>
<td>HERMANSKY-PUDLAK SYNDROME 2</td>
</tr>
<tr>
<td>P53677</td>
<td>AP3M2</td>
</tr>
<tr>
<td>P53677</td>
<td>AP-3 complex subunit mu-2</td>
</tr>
</tbody>
</table>

Table 2. Statistics of dictionaries used in the experiments

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Number of IDs</th>
<th>Number of names</th>
<th>Number of names per ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>14893</td>
<td>205909</td>
<td>13.8</td>
</tr>
<tr>
<td>Mouse</td>
<td>11753</td>
<td>111702</td>
<td>9.5</td>
</tr>
<tr>
<td>E. coli</td>
<td>4875</td>
<td>37095</td>
<td>7.6</td>
</tr>
<tr>
<td>Yeast</td>
<td>5914</td>
<td>59020</td>
<td>10.0</td>
</tr>
<tr>
<td>Drosophila</td>
<td>2376</td>
<td>30891</td>
<td>13.0</td>
</tr>
<tr>
<td>BioCreAtIvE II</td>
<td>32975</td>
<td>182996</td>
<td>5.5</td>
</tr>
<tr>
<td>Overall</td>
<td>72786</td>
<td>627613</td>
<td>8.6</td>
</tr>
</tbody>
</table>

3Since the dictionaries for E.coli, Yeast and Drosophila were small, we were able to carry out experiments without this reduction process, but the performance differences were very small. The recall scores at rank 1 were 51.5, 60.1 and 63.6%, respectively (see Table 4 for comparison).
Furthermore, we limited the maximum number of names used in the training to 32,000.

4.1 Features
Logistic regression modelling allows us to incorporate a variety of features. Since the performance of machine learning heavily depends on how to represent the samples, it is important to use features that can well characterize a string pair. In other words, the features should be able to capture the similarity between a variety of variants (e.g. orthographical, morphological, syntactical, modifiers) while highlighting the difference between those terms which are not synonymous.

In our method, we use the following types of features.

4.1.1 Character bigrams We use the common character bigrams between the two input strings as features. For example, ‘IL2’ consists of two bigrams (‘IL’ and ‘L2’), and ‘IL2R’ consists of three bigrams (‘IL’, ‘L2’ and ‘2R’). We use the shared bigrams (‘IL’ and ‘L2’) as binary features. We also use the value of character bigram similarity as a real-valued feature. The similarity between orthographical, morphological and syntactic variants is expected to be captured by these character n-gram based features.

4.1.2 Prefix/suffix The prefix features focus on the prefixes of the strings. Up to three characters are extracted from the beginning of each string, and the combination of them are used as features. Similarly, the suffix features focus on the suffixes of the strings.

4.1.3 Sharing the same number The numbers in the names often convey important information. For example, the string pair ‘GATA binding protein 2’ and ‘GATA binding protein 5’ shares many characters, but the difference of ‘2’ and ‘5’ indicates that they belong to different IDs. We use a binary feature which indicates whether the strings contain the same number or not.

4.1.4 Acronym We define a feature which can capture the possibility that one string is an acronym of the other. More specifically, this binary feature indicates whether all the characters in the shorter string are included in the longer string in the same order.

4.1.5 Common tokens In addition to the character-level features described above, we use token-level features. We first tokenize each string using white spaces and some predefined delimiters (‘-’, '/', '(', ')', '[', ']' and ','). We then generate the intersection of the two token sets as features. For example, we get ‘GATA’, ‘binding’ and ‘protein’ as the common tokens from the string pair ‘GATA binding protein 5’ and ‘GATA binding protein factor 5’.

4.1.6 Different tokens We also use the symmetrical difference between the two token sets. For the above example, we generate ‘protein’ and ‘factor’ as the different tokens. This feature type is expected to capture tokens which are not important in conveying the concept of a term.

4.1.7 SoftTFIDF One of the merits of using machine learning is that we can incorporate information from a different similarity measure. We use the value of SoftTFIDF similarity as a real-valued feature.

5 EXPERIMENTS
We present experiments comparing our similarity measure to other approaches that use soft string matching.

5.1 Existing soft-matching methods
For the purpose of performance comparison between our proposed method and existing soft-matching methods, we used the following four existing methods.

5.1.1 Levenshtein distance This is also known as the uniform cost edit distance. The Levenshtein distance between two strings $s_1$ and $s_2$ is defined as the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion or substitution of a single character. Note that this measure is defined as distance, but a distance value is easily convertible to a similarity value in an obvious way.

5.1.2 Jaro-Winkler measure The Jaro measure between $s_1$ and $s_2$ is defined as:

$$Jaro(s_1, s_2) = \frac{1}{3} \left( \frac{|s_1|}{|s_1|} + \frac{|s_2|}{|s_2|} + \frac{|\text{shared}(s_1, s_2)|}{|s_1|} \right),$$

where $|s_1|$ and $|s_2|$ are the lengths of $s_1$ and $s_2$, respectively. $|\text{shared}(s_1, s_2)|$ is the number of ‘matching’ characters in $s_1$, where a character in $s_1$ is considered matching if there is the same character in $s_2$ and they are not farther than $\min(|s_1|, |s_2|)/2$. $|\text{shared}(s_1, s_2)|$ is defined analogously. $T_{s_1, s_2}$ is the number of character positions at which the character from $s_1$ and the one from $s_2$ are different.

Let $p_0$ be the number of common prefix characters between $s_1$ and $s_2$. The Jaro–Winkler measure is

$$Jaro – Winkler(s_1, s_2) = Jaro(s_1, s_2) + \frac{p_0}{10}(1 – Jaro(s_1, s_2)),$$

where $p = \max(p_0, 4)$.

5.1.3 Hidden Markov model-based approach (Smith et al., 2003) The similarity value is defined as the probability of generating matching operations with a hidden Markov model. The cost values of matching operations are optimized using synonymous pairs of strings and a forward–backward algorithm.

We used their source code available at their ftp site for the implementation. The model has several meta-parameters. We used the best setting for the meta-parameters reported in Yeganva et al. (2004), i.e. the number of states was set to three, and the model was forced to be symmetrical.

To compute the similarity score from the output of the HMM, we used the following equation as in Smith et al. (2003).

$$\text{score}(s_2, s_1) = \log_{10} \Pr(s_1, s_2) – \log_{10} \Pr(\text{null}, s_2)$$

5.1.4 SoftTFIDF We follow the definition of the SoftTFIDF similarity given in Cohen and Minkov (2006). Each token $t_i$ in string $s$ is given a weight value $w(t_i, s)$, which is computed as $\log(1 + TF) \times \log(IDF)$, where $TF$ is the frequency of the word in the dictionary and $IDF$ is the inverse of the fraction of names in the dictionary that contain that word.

The SoftTFIDF similarity is given by

\[ \text{SoftTFIDF}(s_1, s_2) = \frac{\sum_t w(t_i, s_1) w(t_j, s_2) \text{sim}(t_i, t_j)}{\sqrt{\sum_t w(t_i, s_1)^2} \sqrt{\sum_t w(t_j, s_2)^2}}, \]

where \( \text{sim}(t_i, t_j) \) is the Jaro-Winkler measure between tokens \( t_i \) and \( t_j \) if the measure is equal or greater than 0.9, or 0 otherwise.

### 5.2 Dictionary lookup evaluation with held-out entries

What we evaluate in our experiments is how accurately the various similarity measures enable us to map gene/protein names with the correct IDs. The first set of experiments uses held-out entries from the dictionary as the evaluation data. For each species-specific dictionary that we derived from BioThesaurus, we randomly select 1000 entries and remove them from the dictionary. These removed entries are kept as the evaluation set. We tune (or learn) string similarity using the remaining entries in the dictionary. Finally, we evaluate the performance using the evaluation set.

The gene/protein name of each entry in the evaluation set is matched against the entries in the corresponding dictionary using the similarity measure learned on the dictionary. We define the score for each ID as the maximum similarity value for the gene/protein names that belong to the ID. The system then ranks the IDs according to their scores.

Table 3 shows an example of a ranked list of IDs for the input ‘Acetylating enzyme for N-terminal of ribosomal protein S5’, which is matched against the \( E. \text{coli} \) dictionary. The correct ID for this protein name is ‘P0A948’, which has been ranked second by the system for some (here unspecified) similarity measure.

Once we obtain the ranked ID lists according to each similarity measure for all the samples in the evaluation set, we can evaluate the recall score for retrieving correct IDs at each rank, which is given by \( \frac{M_i}{N} \), where \( N \) is the number of samples in the evaluation set, and \( M_i \) is the number of correct IDs included in the top \( i \) IDs output by the system.

Figures 1–5 compare the performance of the five different methods for each species-specific set. The \( x \)-axis gives the ranks, and the \( y \)-axis is the recall value achieved at each rank.

As expected, the simple similarity measures based on the Levenshtein distance and the Jaro-Winkler metric were confirmed as not good as the learnable similarity measures. The relative performance varies depending on the species-specific dictionary, but in most cases SoftTFIDF performed slightly better than hidden Markov models. The logistic regression method gave the best results with large margins of more than 10% except for the \( E. \text{coli} \) data.

We should note that the absolute performance reported in Figures 1–5 does not necessarily reflect the difficulty of mapping gene/protein names in text to their IDs. For example, the high number of polysemous terms in a dictionary is a major cause of performance deterioration in these experiments, but it does not necessarily mean that the actual mapping task is difficult because such polysemous terms may not often appear in text.

One may be interested in how general each similarity measure is. Our preliminary experiments indicated that the logistic regression-based similarity measure is highly tuned to the training dictionary. In other words, the similarity measure trained for one species is not very useful for a different species. One possible way of creating a more ‘general’ similarity measure is to use a dictionary containing entries from multiple species in training, which, however, entails an increased cost for training.

### Table 3. Ranked list of IDs for the input ‘Acetylating enzyme for N-terminal of ribosomal protein S5’ being matched against the \( E. \text{coli} \) dictionary

<table>
<thead>
<tr>
<th>Rank</th>
<th>ID</th>
<th>Score</th>
<th>Gene/protein name</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P0A7W1</td>
<td>0.969</td>
<td>Ribosomal protein S5</td>
<td>0.969</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3OS ribosomal protein S5</td>
<td>0.896</td>
</tr>
<tr>
<td>2</td>
<td>P0A948</td>
<td>0.824</td>
<td>Ribosomal-protein-S5-alanine acetylase</td>
<td>0.824</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ribosomal-protein-alanine acetyltransferase</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ribosomal-protein-alanine acetyltransferase</td>
<td>0.064</td>
</tr>
<tr>
<td>3</td>
<td>P0A944</td>
<td>0.125</td>
<td>Ribosomal-protein-alanine</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N-acetyltransferase rimI</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ribosomal-protein-alanine acetyltransferase</td>
<td>0.064</td>
</tr>
</tbody>
</table>

The correct ID for this protein name is ‘P0A948’

\(^{5}\)For all experiments, we applied very basic normalization (conversion of capital letters to lower case and hyphens to spaces) to the strings prior the use of similarity measures. This normalization has been shown to have little side effect (Cohen et al., 2002).
5.3 Dictionary lookup evaluation with gene/protein names appearing in text

The training data set provided by the BioCreAtIvE II (2006) gene normalization task contains human gene/protein name snippets from MEDLINE abstracts and their EntrezGene identifiers. For instance, the data provide a snippet ‘Ah (dioxin) receptor’ and its EntrezGene ID ‘196’ for the sentence “The Ah (dioxin) receptor binds a number of widely disseminated…” The data also provide a dictionary, which we used for training the similarity measures.

By using these data, we can conduct dictionary lookup experiments similar to the ones presented in the previous section. The difference is that this evaluation uses actual gene/protein names attested in text rather than the gene/protein names held out from the dictionary. This setting could be seen as the situation where we have a tagger that can perfectly identify gene/protein names in text. Thus, the difficulties of name tagging are not considered here.

Note that we do not consider the problem of ambiguity either. In the actual BioCreAtIvE II (2006) gene normalization task, one would need to perform disambiguation for the polysemous names using the context in which the name appears. This setting could be seen as the situation where we have a tagger that can perfectly identify gene/protein names in text. Thus, the difficulties of name tagging are not considered here.

5.4 Contribution of each feature type

Table 4 shows how each feature type contributed to the dictionary-lookup performance. The first row shows the recall
values at rank 1 achieved when all feature types were used. The other rows show the performance achieved when one of the feature types was unused. The amount of contribution from each feature type varies depending on the dictionary, but bigrams, prefixes and suffixes were, overall, most influential.

6 CONCLUSION
We have described a method for learning a string similarity measure using a logistic regression model and a gene/protein name dictionary. We evaluated our method with several dictionary look-up tasks. Experimental results show that our logistic regression-based method outperforms existing soft-matching methods.

The simplest application of our similarity measure would be in a user interface for a database where one can search for the ID of a gene/protein of interest by using its name. Our technique is also applicable in information extraction systems to aid mapping from text strings to canonical entries. Further application can be found in areas such as ontology construction and merging; aids to authors (mapping of author usage to preferred usage); and term classification (Sparic et al., 2005) and term clustering which can be used for advanced information retrieval/extraction applications.

ACKNOWLEDGEMENTS
We thank P. Cotter, Y. Sasaki for many valuable comments and discussions, and also the reviewers. Our thanks to the Rehholz Text Mining Group at EMBL-EBI, Hinxton, for domain expertise related to bio-resources. This research was supported by EC project BootStrep FP6-028099 (www.bootstrep.org). The UK National Centre for Text Mining is sponsored by the JISC/BBSRC/EPsrc.

Conflict of Interest: None declared.

REFERENCES


