Mining semantic networks of bioinformatics e-resources from the literature

Mining Semantic Networks of Bioinformatics e-Resources from the Literature

Hammad Afzal[1,2], James Eales[1], Robert Stevens[1], Goran Nenadic[1]

[1]School of Computer Science, University of Manchester,
Oxford Road, Manchester, M13 9PL, UK
[2]DERI, Unit for Natural Language Processing,
National University of Ireland, Galway, Ireland
{Hammad.Afzal@postgrad., James.Eales@, Robert.Stevens@,
G.Nenadic@}manchester.ac.uk

Abstract. There have been a number of recent efforts (e.g. BioCatalogue, BioMOBY, etc.) to systematically catalogue bioinformatics tools, services and datasets. These efforts mostly rely on manual curation and are unable to cope with the huge influx of various electronic resources, which consequently result in their unavailability to the community. We present a text mining approach that utilizes the literature to extract and semantically profile bioinformatics resources. Our method identifies the mentions of resources in the literature and assigns a set of co-occurring terminological and ontological entities (descriptors) to represent them. Since such representations can be extremely sparse, we use kernel metrics based on lexical term/descriptor similarities to identify semantically related resources. Resources are then either clustered or linked into a network, providing the users (bioinformaticians and service/tool crawlers) with a possibility to explore tools, services and datasets based on their relatedness, thus potentially improving the resource discovery process.

Keywords: bioinformatics services, service description, text mining, networks, kernel similarity.

1 Introduction

The rapid increase in the amount of bioinformatics data produced in recent years has resulted in the huge influx of bioinformatics electronic resources (e-resources), such as online-databases [1], data-analysis tools, Web services [2] etc. Discovering such resources became a major bottleneck in bioinformatics: in order to effectively utilize them, e-resources need to be organised and their functionalities semantically described. A number of community wide efforts such as BioCatalogue [3] and BioMoby [4] have been initiated to systematically catalogue the "resourceome". By annotating services using keywords and ontological concepts, such catalogues facilitate access to both bioinformaticians and Semantic Web crawlers and agents that can orchestrate the use of such resources. However, the annotation process depends on a typically slow manual curation process that hinders the growth of such curated

resources to keep pace with the very field they attempt to catalogue. For instance, the number of registered services in BioCatalogue (there are 1,084 of them[1]) is still small compared with the total number of Web services available online: it is estimated that there are ~3500 life science Web services in Taverna alone [3, 5]. This fact calls for the development of semi-automatic methods for resource annotation and their cataloguing in order to maximise the utility of e-resources by making them widely available to the community.

One of the key aims of providing bioinformatics resources with semantic descriptions is to improve resource discovery. Semantically-described resources can not only be searched, browsed and discovered by using keyword-based queries (for instance, via their names or task descriptions), but also on the basis of the semantic relatedness of their functionalities. For example, BioCatalogue descriptions refer to *similar services* (see Fig. 1) so that the users can identify related tools.

Fig. 1. A snapshot of a Web service description taken from BioCatalogue[2]

When manually assigned annotation tags and/or related services are not available, we hypothesise that automated approaches could be used to improve the discovery process. These include building networks and clusters of similar resources. For example, a user can search for a Web service that corresponds to a particular input, output or operation performed. If, however, the retrieved services do not fulfil the exact requirement, the user may be interested in exploring similar services (for example, with more generic/specific input/output, but still with a related functionality), which can be identified by browsing a Web service network or by exploring clusters of related services. Traditionally, similar or related services have been identified by using *lexical comparisons* of their names and names of their parameters (input/output) and operations. This process has been further improved by *concept-based comparisons* using domain ontologies that have been used to annotate the resources (as in myGrid [6] and BioCatalogue).

[1] Statistics collected on 10th Oct, 2009.

[2] http://www.biocatalogue.org/services/2048-wsblastpgpservice_414364

In our previous work, we have shown that the vast amounts of scientific literature related to bioinformatics resources can be tapped in order to automatically extract their key semantic functional features [7, 8]. In this paper we propose a methodology to build and explore clusters and semantic networks of bioinformatics resources, which can help to identify related resources on the basis of their similarity as well as by their semantic relatedness. In order to measure the semantic relatedness between the resources, we have designed a kernel-based similarity approach that uses lexical and semantic properties of resource mentions as extracted from the literature.

2 Methodology

The overall methodology adopted in the work presented here is based on the concepts of bioinformatics resources, semantic resource descriptors, and kernel/similarity functions, which are explained below.

Bioinformatics resources represent the list of e-resources which are used by bioinformaticians while performing in-silico experiments [9, 10]. We have focused on the four major classes: *Algorithms*[3], *Applications*, *Data* and *Data Resources*. These have been engineered from the myGrid ontology. Table 1 shows example resource instances belonging to these classes. In our previous work, we have described a set of text mining tools that can be used to efficiently identify, classify and extract mentions of these resources in the literature [7]. The method is based on key terminological heads assigned to each of the semantic classes (e.g. *alignment* and *method* are "linked" to *Algorithms*, while *sequence* and *record* point to a *Data* entity) and specific lexico-syntactic patterns (enumerations, coordination, etc.).

Table 1. Examples of semantic classes and their instances

Semantic class	Example instances
Algorithm	SigCalc *algorithm*, CHAOS local *alignment*, SNP *analysis*, KEGG Genome-based *approach*, GeneMark *method*, K-fold cross validation *procedure*
Application	PreBIND Searcher *program*, Apollo2Go *Web Service*, FLIP *application*, Apollo Genome Annotation curation *tool*, GenePix *software*, Pegasys *system*
Data	GeneBank *record*, Genome Microbial CoDing *sequences*, Drug Data *report*
Data resource	PIR Protein Information *Resource*, BIND *database*, TIGR *dataset*, BioMOBY Public Code *repository*

Semantic resource descriptors are the key terminological phrases used in the existing textual descriptions of bioinformatics resources, as given by various providers such as BioCatalogue, BioMoby, EBI[4], etc. These descriptors refer to

[3] Note that, to aid simplicity and uniformity, we consider *Algorithms* as e-resources.
[4] http://www.ebi.ac.uk/Tools/webservices/

concepts and specific roles (e.g. input/output parameters, etc) and have been frequently used in the existing descriptions. For example, frequent descriptors are *gene expression*, *phylogenetic tree*, *microarray experiment*, *hierarchical clustering*, *amino acid sequence*, *motif*, etc. We use such descriptors to profile a given resource and/or to link it to a domain ontology.

We have used two sources to build a dictionary of bioinformatics resource descriptors. The first source is the list of terms collected from the bioinformatics ontology used in the myGrid project [11]. This list contains 443 terms describing concepts in *informatics* (the key concepts of data, data structures, databases and metadata); *bioinformatics* (domain-specific data sources e.g. model organism sequencing databases, and domain-specific algorithms for searching and analysing data e.g. a *sequence alignment algorithm*); *molecular biology* (higher level concepts used to describe bioinformatics data types, used as inputs and outputs in services e.g. *protein sequence*, *nucleic acid sequence*); and *tasks* (generic tasks a service operation can perform e.g. *retrieving*, *displaying*, *aligning*). The second source includes automatically extracted terms (recognised by the TerMine[5] service) and frequent noun phrases obtained from existing descriptions of bioinformatics Web resources available from BioCatalogue.

For each bioinformatics resource that can be identified in the literature, we build its *semantic profile* by harvesting all descriptors that co-occur with the resource in the same sentence in a given coprus (see Fig. 2 for an example). These profiles are then used to establish semantic similarities between resources by comparing the descriptors (used as features) that have been assigned to them.

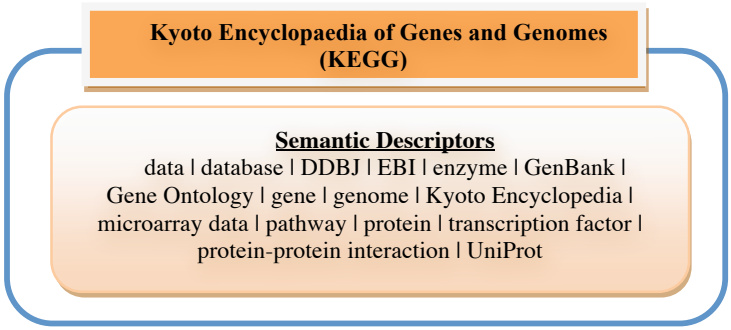

Fig. 2. Semantic Resource Descriptors for the Kyoto Encyclopaedia of Genes and Genomes

Since service representations using descriptors can be extremely sparse, we use **kernel metrics** based on term/descriptor similarities to identify semantically related resources. The main aim is to enhance the comparison process by incorporating lexical and contextual properties of *descriptors* retrieved from the literature. This approach is inherent to our method, as descriptors (used as features for the resources) have been retrieved from sentences that are related to resources. Various similarity kernels can be used for comparisons (e.g. bag-of-words kernels [12, 13], string

[5] http://www.nactem.ac.uk/software/termine/

kernels [14], etc.). Here we have considered three approaches which are described below.

- **Method 1: lexical comparison of resource names**. This is a simple similarity function that relies on lexical profiles of resource names. The *lexical profile* of a term comprises all possible linear combinations of word-level substrings present in that term [15]. For example, the lexical profile of term '*protein sequence alignment'* comprises the following terms *protein*, *sequence*, *alignment*, *protein sequence, sequence alignment*, *protein sequence alignment*. In this method, the similarity between two resources is then calculated as a similarity between lexical profiles of their names. Formally, let LP(s_1) and LP(s_2) be lexical profiles (represented as vectors) of names of resources s_1 and s_2. Then the similarity function is defined as:

$$\mathrm{Sim}_1(s_1,\ s_2) = \frac{\mathrm{LP}(s_1) \cdot \mathrm{LP}(s_2)}{|\ \mathrm{LP}(s_1)\ | \cdot |\ \mathrm{LP}(s_2)|} \quad (1)$$

- **Method 2: shared descriptors**. Another option is to use the standard bag-of-descriptors kernel, where each resource is represented as a bag of its descriptors and the similarity is based on exact matches between descriptors. This kernel compares the resources using the inner product that measures the degree of descriptor sharing:

$$\mathrm{Sim}_2(s_1, s_2) = s_1 \cdot s_2 \quad (2)$$

  where s_1 and s_2 are vectors that represent the semantic descriptors assigned to the resources being compared. Alternatively, cosine similarity can be used if we use the frequency of the occurrence of the semantic descriptors (not presented here).

- **Method 3: lexical similarity of shared descriptors**. A further option for a kernel function is to use the relatedness between descriptors to measure the similarity between the resources. The main motivation behind this approach is that resources can share related but not exactly the same descriptors. We therefore suggest using a kernel that takes into account *descriptor smoothing* by incorporating a similarity measure between descriptors themselves in the kernel function that calculates similarity between resources. Formally, let $S = \{s_1, ..., s_k\}$ be the set of e-resources whose descriptions have been collected from the literature. Let $D = \{d_1, ..., d_m\}$ be the set of all descriptors, where m is the total number of descriptors. In order to measure similarity between two resources, we first build a similarity matrix $\boldsymbol{A}$ (m x m), where each element a_{ij} corresponds to the similarity between descriptors d_i and d_j. Then, the similarity between two resources s_1 and s_2 is calculated as:

$$\mathrm{Sim}_3(s_1, s_2) = s_1 \cdot A \cdot s_2 \quad (3)$$

  In the experiments reported below, the smoothing is done by calculating the cosine similarity between the lexical profiles of the descriptors (analogously to (1)).

3 Experiments and Discussion

Here we demonstrate the development of networks of related resources by using each of the three methods stated above. The networks are visualised as weighted, undirected graphs where nodes are resources and edges represent relatedness between them. This *relatedness* is estimated using the similarity functions, where the weight of an edge represents the strength of the relationship between the two connected nodes (see Section 3.2). We also investigate different methods of exploring and visualising our similarity matrices; specifically we use hierarchical clustering dendrograms and heatmap visualisations.

3.1 Data

Table 2 gives the number of bioinformatics resources that were identified in a corpus of 2,691 full-text articles published by the journal BMC Bioinformatics. The details of the extraction process are presented in [7].

Table 2. The statistics of Bioinformatics e-resources found in the BMC Bioinformatics corpus

Semantic Class	Total # of instances	Average # of descriptors
Algorithm	5,722	9
Application	2,076	8
Data	2,662	15
Data Resource	1,992	10

Each of the e-resources has been assigned a set of associated descriptors (11 descriptors on average; see Table 2 for details for the specific classes). As can be expected, single word descriptors appeared more frequently in the corpus. Table 3 lists the most frequent single word, two- and three-word descriptors.

Table 3. The most frequent single-word, two-word and three-word descriptors

Single word descriptors	Two-word descriptors	Three-word descriptors
gene: 13,585	*gene expression*: 1,147	*protein-protein interaction*: 308
method: 8,203	*secondary structure*: 887	*multiple sequence alignment*: 295
protein: 6,417	*protein sequence*: 780	*gene expression data*: 262
sequence: 5,991	*protein structure*: 574	*amino acid sequence*: 257
analysis: 4,287	*microarray experiment*: 488	*Smith-Waterman algorithm*: 48

3.2 Exploration of Semantic Networks

Here we assess the utility of resource descriptors for semantic profiling of bioinformatics resources. We do this by exploring our hypothesis that bioinformatics resources can be semantically linked via resource descriptions. For this, we have manually identified a sample of 18 resources that are commonly used in bioinformatics (see Table 4). Each of these has occurred in more than 120 sentences in our corpus. The sample contains resources from all four semantic classes of resource. The results have been generated using the three methods for deriving semantic relatedness between resources as described above.

Table 4. A sample of resources used for exploration

Resource Name	Number of sentences	Resource Class
Gene ontology (GO)	6757	Data resource
Support vector machine (SVM)	2456	Algorithm
Protein data bank (PDB)	904	Data resource
Hidden Markov model (HMM)	602	Algorithm
Principal components analysis (PCA)	599	Algorithm
Position-specific scoring matrix (PSSM)	457	Algorithm
Self organising map (SOM)	305	Algorithm
Medical subject headings (MeSH)	261	Data resource
Neural network	256	Algorithm
Markov chain Monte Carlo (MCMC)	252	Algorithm
Expression profile	252	Data
Basic local alignment search tool (BLAST)	238	Application
Phylogenetic tree	233	Data
Structural classification of proteins (SCOP)	216	Data resource
Kyoto encyclopaedia of genes and genomes (KEGG)	187	Data resource
Clusters of orthologous groups (COG)	163	Data resource
ChIp-chip data	126	Data
Pairwise alignment	123	Data

Method 1: lexical comparison of resource names. As expected, this method did not yield useful results as very little similarity was found between resource names. This suggests that surface-level lexical information originating from resource names is not sufficient to develop semantic networks of resources.

Method 2: shared descriptors. We derived mutual similarity scores for the 18 resources, with a mean of 0.34 and a standard deviation of 0.09. This method identified significant relatedness between many resources (see Fig. 3 for a heat-map). Clearly, the addition of descriptors improved our ability to derive a measure of semantic similarity between resources whose names are lexically disparate. Although it is difficult to define any clear semantic relationships from these data, it is noticeable

that *ChIp-chip data* has specific properties that are not commonly matched by others in the sample, (manifested as a line of light yellow on the heat-map, see Fig. 3).

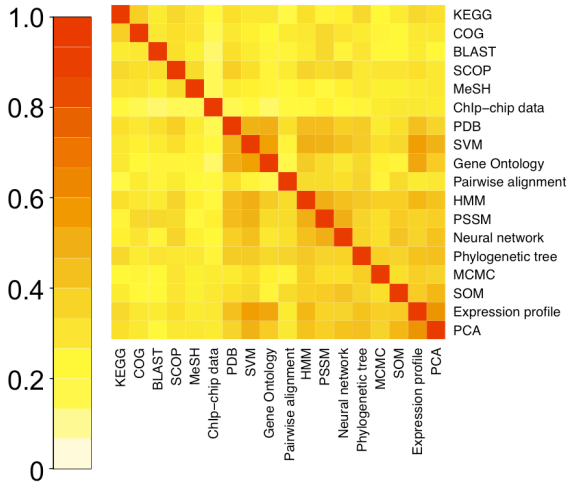

Fig. 3. Heatmap representation of the matrix of shared descriptor similarity scores between resources (method 2). Values vary from 1.0 (red) to 0.00 (white), see legend. Heatmap generated by R function, 'heatmap' [16].

To further highlight the subtle differences and similarities between the resources in the sample, we applied a hierarchical clustering algorithm [17] to the matrix of scores (see the resulting tree in Fig. 4).

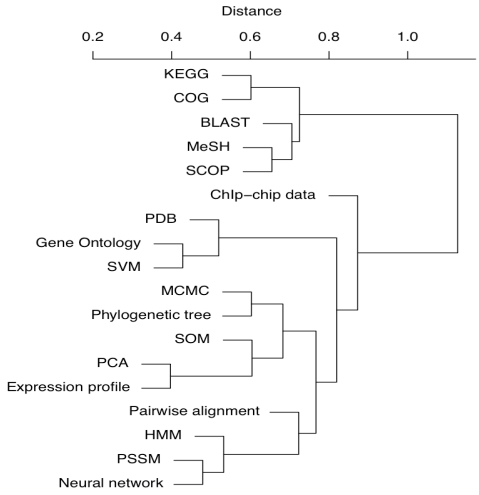

Fig. 4. Hierarchical clustering of e-resources using the shared descriptors similarity matrix (method 2). Distances were calculated as $(1 - Sim_2)$. Ward's minimum variance clustering method [17] was used to cluster the data. The tree was generated using R function 'hclust'.

The tree in Fig. 4 highlights some interesting clusters of the examined resources. Rather than being clustered by resource class, there are semantically important links being identified. For example, *PCA* and *SOM* are important and widely used methods for exploring expression data [18], and these resources form their own cluster. Additionally, there is a link established between *phylogenetic tree* and *MCMC*; *MCMC*, in combination with a Bayesian approach, is a popular method in phylogenetic analysis for the derivation of trees of relationships between sequences [19]. The cluster of *pairwise alignment*, *HMM*, *PSSM* and *neural network* highlights the semantic theme of sequence analysis (*HMM*, *PSSM* and *neural networks* have all been used successfully to analyse pairwise and multiple sequence alignments). *KEGG*, *BLAST*, *COG*, *SCOP* and *MeSH* form their own group, which do not highlight any obvious semantic relationships; a likely reason is that these resources have such broad utility that the specifics of the relationships between them are lost. It is surprising, however, that *GO* and *PDB* did not follow a similar pattern.

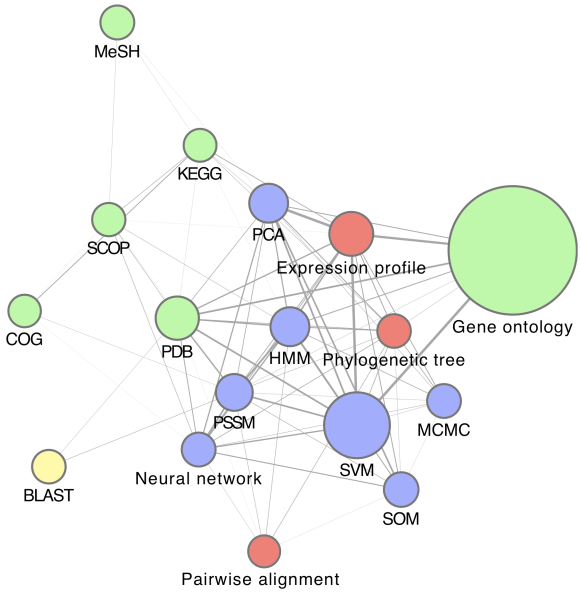

Fig. 5. Semantic network of bioinformatics resources (using method 2 and values shown in Fig. 3). Node size represents frequency in the corpus, edge thickness represents how similar the two connected nodes are. Node colour is determined by the semantic class of the node (red for *Data*, green for *Data resource*, blue for *Algorithm* and yellow for *Application*. The image was generated using Cytoscape[6], the network was laid out using the Cytoscape layout algorithm 'Edge-Weighted Spring Embedded', using the edge weight data in the network.

Even though similarity data alone can identify important semantic links, we further explored the importance of the number and strength of links between resources. In Fig. 5 we present our similarity data as edges in a network connecting each node (representing individual resources) with those that have some similarity to it. Each

[6] Cytoscape, http://www.cytoscape.org/

edge is weighted by the similarity between the resources it connects, so that edges that appear thick represent strong relationships and weak relationships are represented by thin edges. We have removed all edges that have a weight below the median edge weight for the network. Our intention with this was to remove edges that exist due to chance alone and to better highlight the strongest relationships in the network.[7] The strongest links occur between the resources that appear most frequently in the corpus. The strongest link is between *Gene Ontology* and *SVM*, most probably because *SVM* methods have been widely used for protein annotations using GO (see, for example, [26]). Strong links also occur between *PCA* and *expression profile* and *expression profile* and *SVM*, indicating types of algorithms used with specific data types.

Method 3: lexical similarity of shared descriptors. The results of calculations for linking the resources considering the lexical similarities between their descriptors are summarised in figures 6, 7 and 8.

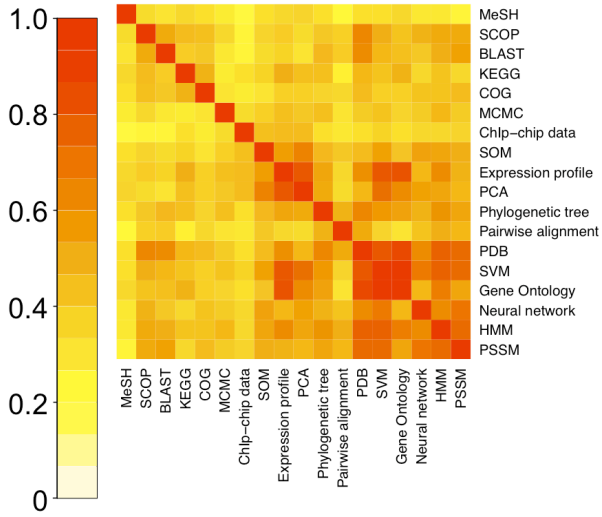

Fig. 6. Heatmap representation of the matrix of lexically smoothed descriptor similarity scores between resources (method 3). Values vary from 1.0 (red) to 0.0 (white), see legend. Heatmap generated by R function, 'heatmap' [16].

Fig. 6 has some similarity with Fig. 3. However, there are clusters of more closely related resources (for example *expression profile*, *SVM* and *Gene Ontology*). All resources again have some similarity to all others, making it more difficult to identify the most important relationships. This indicates that sensible thresholds need to be identified to remove uninteresting links. The similarity scores have a mean of 0.47 and a standard deviation of 0.14, which suggests that the scores from method 3 vary more widely than those from method 2 and thus potentially provide better discrimination.

[7] *ChIp-chip data* is missing from Fig. 5 because all its edges have weights below the median.

Fig. 7 (the hierarchical tree) is similar to Fig. 4 in the sense that some of the clusters are shared between the trees. The cluster of resources with broad implications and uses (*MeSH*, *BLAST, SCOP, KEGG* and *COG*) in particular is still present. However, some new interesting clusters have emerged: for example, the data resources *phylogenetic tree* and *pairwise alignment* have been clustered together, both of which are common data forms in sequence analysis.

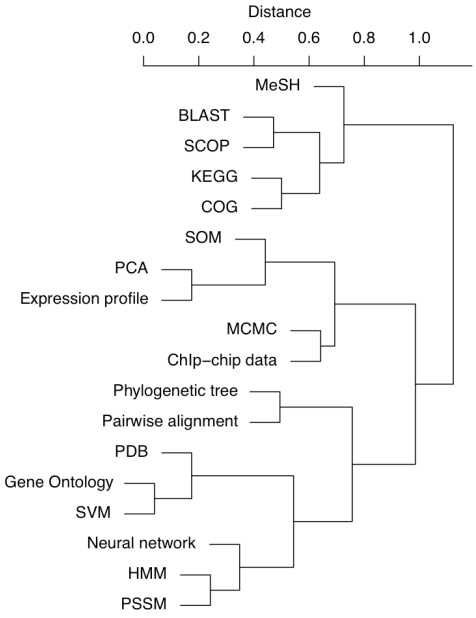

Fig. 7. Hierarchical clustering of e-resources using the lexically smoothed similarity matrix (method 3). Distances were calculated as $(1 - \mathrm{Sim}_3)$. Ward's minimum variance clustering method [17] was used to cluster the data. Tree generated using R function 'hclust' [16].

The network[8] given in Fig. 8 presents the strongest clustering of resources based on their class. Although the *Data* nodes (represented as red) are not strongly linked to each other, the *Data Resource* nodes (green) are all clustered together. There is also a similar pattern with the *Algorithm* nodes (blue). The strongest edge weights again occur between resources that appear most frequently in the corpus, suggesting that frequency normalisation may be needed to reduce this impact. *Gene Ontology*, in particular, is linked to all other resources, and that is primarily a product of its ubiquity in the literature and therefore the tendency for many descriptors and resources to be linked to it.

[8] We have again removed the edges with a weight below the median edge weight. This has caused the removal of the *MeSH* node from the network. This could be due to *MeSH* being the only resource strongly related to literature resources.

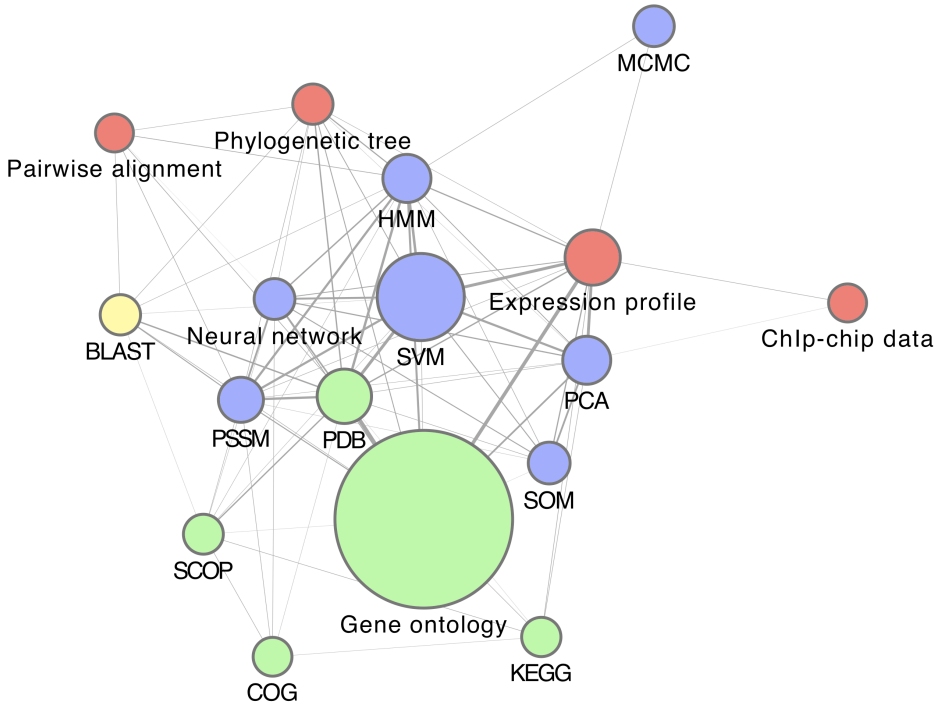

Fig. 8. Semantic network of bioinformatics resources (using method 3 and values shown in Fig. 6). Node size represents frequency in the corpus, edge thickness / weight represents how similar the 2 connected nodes are. Node colour is determined by the semantic class of the node, red for *Data*, green for *Data resource*, blue for *Algorithm* and yellow for *Application*. The image was generated using *Cytoscape* (see Fig. 6 for details).

By further analysing the associated semantic profiles, we can see that significant relatedness between resources typically originates from sharing a number of very generic descriptors, in particular single-word ones (see Table 3). Many of these have a generic nature (e.g. *method*, *analysis*, *gene*, etc.). This problem, however, can be resolved by either filtering generic concepts from the descriptors using stop-words or by using a *tf*idf*-like weights [20] assigned to descriptors (considering the frequency of descriptors appearing in profiles of different resources). Selecting and varying the threshold for the edge weight in our network representations can also discard unwanted weak links.

4 Related Work

Most of the efforts in the domain of Semantic Web for the Life Sciences have been focused on data annotation (e.g. a number of protein function databases), using both manual and automated approaches. Recently, these efforts have been extended to semantic description of services and tools that are used to analyse, visualise and explore such data.

In addition to manual annotation, Semantic Web technologies have been applied for the description of Web services. These approaches include descriptions of Web service functionalities as well as the meta-data information about their inputs and outputs. Most of the suggested approaches rely on the data available in WSDL files. For example, Lerman and colleagues [21] presented work on automatic labelling of inputs and outputs of Web services using meta-data based classification relying on terms extracted from the associated WSDL files. The underlying heuristic behind the meta-data based classification is that similar data types tend to be named by similar names and/or belong to operations or messages that are similarly named. Similarly, Hess and Kushmerick [22] used machine learning to classify Web services using information given in WSDL files of the services which include port types, operations and parameters along with any documentation available about the Web service. Information in a WSDL file is treated as "normal" text, and the problem of Web service and its metadata classification is addressed as a text classification problem.

Carman and Knoblock, on the other hand, reported on invoking new/unknown services and comparing the data they produce with that of known services, and then use the meta-data associated with the known services to add annotations to the unknown resources [23]. Belhajjame and colleagues [24] used known annotations of parameters belonging to components in a workflow to infer the unknown annotations of other parameters (in other components). Here, semantic information of operation parameters is inferred based on their connections to other (annotated) components within existing tried-and-tested workflows. Apart from deriving new annotations, this method can inspect the parameter compatibility in workflows and can also highlight conflicting parameter annotations.

There have been some efforts to improve the service discovery process. Dong et al. [25], for example, used clustering-based approach in which parameters of service operations are grouped into meaningful concepts, which are then used to find similar service operations based on similar parameters. However, this method provides only a limited solution and is unable to provide comprehensive service discovery based on the underlying semantics provided by services. Employing Semantic Web approaches such as ontological annotations could improve this approach [11].

5 Conclusions

In this paper we proposed and explored a literature-based methodology for building clusters and semantic networks of functionally related e-resources in bioinformatics. The main motivation is to facilitate the resource discovery approaches, which would improve the availability and utility of these resources to the community. The methodology revolves around terminological units (semantic descriptors) that are frequently used by bioinformatics resource providers to semantically describe the resources. The semantic descriptors have been automatically compiled and each e-resource has been assigned a set of descriptors co-occurring with the given e-resource in a full-text article corpus.

In order to establish similarity between resources, their profiles are compared using three levels of service representations: the lexical similarity between the resource names (method 1); the similarity calculated on the basis of shared semantic

descriptors (method 2), and the same similarity smoothed by considering lexically similar descriptors. As expected, the first method failed to capture any significant links between resources as it relied solely on the surface level clues originating from the names of resources. The second approach performed significantly better and was able to identify interesting clustering patterns between the resources which did not have any lexical resemblance. At the third level, in contrast to considering the exact match between resource descriptors, we devised a descriptor-based kernel matrix, which incorporated the approximate lexical similarities between the descriptors (using their lexical profiles). The approximate similarities helped in linking the resources that shared the descriptors which were not exactly the same, but were related. An interesting pattern emerged whilst experimenting with this metric, whereby resources would cluster together based on their class (i.e. the resources which belonged to the same class (such as *Algorithm*, *Data Resource*, etc.) tended to appear closer in the network). Method 2 revealed some interesting functional links (linking data types and algorithms). It remains an open question as to which of these clustering patterns is most useful for semantic resource discovery.

The work presented here demonstrates the potential of simple kernel methods (using lexical profiles) built to model relatedness between resource descriptors. We anticipate that further work will be required to identify the most relevant weights for semantic descriptors to counter-balance the impact of frequent (and less informative) features. Other kernels (such as contextual and distributional similarities, WordNet-based similarities, string kernels etc) need to be explored and could provide better resolution of the complex interrelationships between resources.

Acknowledgments. This work was partially supported by the UK Biotechnology and Biological Science Research Council (BBSRC) via the "BioCatalogue" project and by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2) awarded to HA. JE is funded by the e-LICO project (EU Grant agreement number 231519).

References

1. Nucleic Acids Research, DB Issue. Volume 37 (January, 2009)
2. Nucleic Acids Research, Web Server Issue. Volume 37 (July, 2009)
3. Goble, C A., Belhajjame, K., Tanoh, F., Bhagat, J., Wolstencroft, K., Stevens, R., Nzuobontane, E., McWilliam, H., Laurent, T. & Lopez, R.: BioCatalogue: A Curated Web Service Registry For The Life Science Community. In 3rd International Biocuration Conference, Berlin Germany (2009)
4. Wilkinson, MD., Links, M.: BioMOBY: an open source biological web services proposal. Briefings in Bioinformatics. 3:331–41 (2002)
5. Oinn, T., Greenwood, M., Addis, M., Alpdemir, N., Wroe, C., et al.: Taverna: lessons in creating a workflow environment for the life sciences: Research Articles. Concurr. Comput. : Pract. Exper. 18(10): pp. 1067-1100, (2006)
6. Stevens, R., Robinson, A., Goble, C.: myGrid: personalised bioinformatics on the information grid. In Bioinformatics (ISMB Supplement), pp. 302-304 (2003)
7. Afzal, H., Stevens, R., Nenadic, G.: Mining Semantic Descriptions of Bioinformatics Web Resources from the Literature. In Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications. Heraklion, Crete, Greece, Springer-Verlag, pp. 535-549. (2009)

8. Afzal, H., Stevens, R., Nenadic, G.: Towards Semantic Annotation of Bioinformatics Services: Building a Controlled Vocabulary. In Proc. of the Third International Symposium on Semantic Mining in Biomedicine, Turku, Finland, pp. 5–12 (2008)
9. Eales, JM., Pinney, JW., Stevens, RD., Robertson, DL.: Methodology capture: discriminating between the "best" and the rest of community practice. BMC Bioinformatics 9: 359 (2008).
10. Fisher P, Hedeler C, Wolstencroft K, Hulme H, Noyes H, Kemp S, Stevens R, Brass A: A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis. Nucleic Acids Res. 35: pp. 5625–5633 (2007)
11. Wolstencroft, K., Alper, P., Hull, D., Wroe, C., Lord, P.W., Stevens, R.D., Goble, C.A.: The myGrid Ontology: Bioinformatics Service Discovery. International Journal of Bioinformatics Research and Applications 3, pp. 326–340 (2007)
12. Teytaud, O., Jalam, R.: Kernel-based text categorization. In International Joint Conference on Neural Networks (IJCNN'2001), Washington DC (2001).
13. Pahikkala, T., Pyysalo, S., Ginter, F., Boberg, J., Jarvinen, J., et al.: Kernels incorporating word positional information in natural language disambiguation tasks. In Proc. of the Eighteenth International Florida Artificial Intelligence Research Society Conference, Menlo Park, California. AAAI Press (2005).
14. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. J. Mach. Learn. Res. 2 pp. 419-444 (2002).
15. Nenadic, G., Ananiadou, S.: Mining Semantically Related Terms from Biomedical Literature. ACM Transactions on Asian Language Information Processing. pp. 22-43(2006)
16. R Development Core Team: R, A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2009)
17. Romesburg, H.C.: Cluster analysis for researchers. Lulu Press, North Carolina (2004).
18. Belacel N, Wang Q, Cuperlovic-Culf M: Clustering methods for microarray gene expression data. Omics. 2006; 10:pp. 507–531 (2006)
19. Liang, LJ., Weiss, RE., Redelings B, Suchard MA.: Improving phylogenetic analyses by incorporating additional information from genetic sequence databases. Bioinformatics; 25(19):pp. 2530-6 (2009)
20. Salton, G, Buckley C.: Term-weighting approaches in automatic text retrieval, Information Processing and Management, pp. 513—523, (1998)
21. Lerman, K., Plangrasopchok, A., Knoblock, C.: Automatically Labelling the Inputs and Outputs of Web Services. In Proceedings of AAAI-2006, Boston, MA, USA: 149-181 (2006)
22. Hess, A., Kushmerick, N. : Learning to Attach Semantic Metadata to Web Services. In Proc. 2nd International Semantic Web Conference (ISWC2003). Sanibel Island, Florida, USA, Springer Berlin / Heidelberg. 2870/2003: 258-273 (2003).
23. Carman, M. J., Knoblock, C. A.: Learning Semantic Descriptions of Web Information Sources. Twentieth International Joint conference on Artificial Intelligence, Hyderabad India: 1474-1480, (2007).
24. Belhajjame, K., Embury, S. M., Paton, N. W., Stevens, R.: Automatic annotation of Web services based on workflow definitions. ACM Trans. Web 2(2): 1-34 (2007)
25. Dong, X., Halevy, A., Madhavan, J., Nemes, E., Zhang, J.: Similarity search for Web services. In Proceedings of the Thirtieth international conference on Very large data bases - Volume 30. Toronto, Canada, VLDB Endowment: pp. 372-383. (2004)
26. Vinayagam, A., König, R., Moormann, J., Schubert, F., Eils, R., Glatting, KH., Suhai, S..: Applying Support Vector Machines for Gene ontology based gene function prediction, BMC Bioinformatics 2004, 5:116