Databases and ontologies

Babel's tower revisited: a universal resource for cross-referencing across annotation databases

Sorin Drăghici*, Sivakumar Sellamuthu and Purvesh Khatri

Department of Computer Science, Wayne State University, 431 State Hall, Detroit, MI 48202, USA

ABSTRACT

Motivation: Annotation databases are widely used as public repositories of biological knowledge. However, most of these resources have been developed by independent groups which used different designs and different identifiers for the same biological entities. As we show in this article, incoherent name spaces between various databases represent a serious impediment to using the existing annotations at their full potential. Navigating between various such name spaces by mapping IDs from one database to another is a very important issue which is not properly addressed at the moment.

Results: We have developed a web-based resource, Onto-Translate (OT), which effectively addresses this problem. OT is able to map onto each other different types of biological entities from the following annotation databases: Swiss-Prot, TrEMBL, NREF, PIR, Gene Ontology, KEGG, Entrez Gene, GenBank, GenPept, IMAGE, RefSeq, UniGene, OMIM, PDB, Eukaryotic Promoter Database, HUGO Gene Nomenclature Committee and NetAffx. Currently, OT is able to perform 462 types of mappings between 29 different types of IDs from 17 databases concerning 53 organisms. Among these, over 300 types of translations and 15 types of IDs are not currently supported by any other tool or resource. On average, OT is able to correctly map between 96 and 99% of the biological entities provided as input. In terms of speed, sets of ~20 000 IDs can be translated in <30 s, in most cases.

Availability: OT is a part of Onto-Tools, which is freely available at http://vortex.cs.wayne.edu/Projects.html

Contact: sorin@wayne.edu

1 INTRODUCTION

Gene annotations databases are widely used as public repositories of biological knowledge. Understanding the results of almost any molecular biology experiment involves consulting such annotation databases. Our current knowledge is spread out over a number of databases (DBs), such as Entrez Gene (Maglott *et al.*, 2005, UniProt (Bairoch *et al.*, 2005), Protein Data Bank (Berman *et al.*, 2000), RefSeq (Pruitt and Maglott, 2001), RGD, SGD, WormBase and Gene Ontology (GO) (Ashburner *et al.*, 2001) to name just a few. Many such databases support multiple organisms but are specialized on a subset of specific biological entities. For instance, UniProt focuses on proteins, Entrez Gene focuses on genes, EPD focuses on eukaryotic promoters, etc. Other databases aim to provide a wider angle but focus on specific organisms. Examples could include RGD for rat, SGD for yeast, WormBase for *Caenarhabditis elegans*, etc. Obtaining a complete understanding of an experiment usually requires combining information from several such annotation databases. Unique key identifiers (IDs) in the internal structure of each such database represent biological entities such as genes, proteins and mRNAs. Design and implementation restrictions specific to each database ensure that, within each database, the data are consistent, coherent and non-redundant. However, most of these annotation databases have been developed by independent groups which have used completely different designs and completely different sets of key identifiers for the same biological entities. Because of this, the ensemble of such annotation databases, which is the current repository of all our biological knowledge is inconsistent, incoherent and highly redundant.

At the same time, the old-fashioned gene-centric approach of research in life sciences has been all but substituted by more high-throughput approaches involving entire sets of genes, sometimes entire genomes. In many current life science experiments, researchers obtain results identifying many genes that are interesting in a given condition. In order to fully interpret such results, researchers must combine annotations from several different databases which essentially requires mapping tens or hundreds of IDs across all databases involved. If performed manually, this mapping often leads to incomplete and incorrect results, and is time consuming and error prone even for short lists of genes. Even if performed automatically, querying various databases for the same data often yields different results. This represents a very important problem that has not been satisfactorily addressed yet.

2 NAME SPACE ISSUES IN ANNOTATION DATABASES

Identifiers used in different databases often represent different types of biological entities (e.g. genes, ESTs, mRNAs, proteins, etc.). Usually, there is a very clear and biologically meaningful mapping from one such entity to another. For instance, in the simplest case, a gene has a unique DNA sequence, which in turn can be mapped to an mRNA sequence, that is translated into a protein sequence, which perhaps has a known protein structure. However, the problem is further complicated by one-to-many mappings at various levels. For instance, several ESTs can represent the same gene, several alternatively spliced mRNAs can be constructed from the same gene DNA sequence, several structures corresponding to alternative folding patterns or different possible ligands can be associated with the same protein, etc. Specific annotations are available at each level (gene, mRNA, protein, structure, etc.). Given for instance a set of

*To whom correspondence should be addressed.

Table 1. Human gene *XBP1* is represented by six additional distinct identifiers (IDs) in six different databases, as well as by one nucleotide sequence ID, one protein sequence ID and seven different probe IDs on several different Affymetrix arrays

Database	Identifier (ID)
UniGene	Hs.437638
HGNC	12801
Entrez Gene	7494
Swiss-Prot	P17861
ENSEMBL	ENSG00000100219
PharmGKB	PA37400
RefSeq	NM_005080, NP_005071
NetAffx	RC_W90128_s_at (HU35ksubd), 200670_at (HGU 133), 71584_at (HGU95e), M31627_at (HG FL), 39756_g_at (HGU 95av2), 39755_at (HG U95a), g4827057_3p_s_at (U133 X3P)

genes found to be differentially expressed in a specific condition of interest, one wishes to quickly find all known annotations about this set of genes, at all levels: the known GO categories associated with each of these genes, their proteins, the annotations associated with these proteins, etc. This information is currently spread out over many different databases, and each such database uses its own type of IDs. For instance, Table 1 shows eight different IDs used to refer to the same *XBP1* gene in seven different databases, as well as seven different probe IDs used on several Affymetrix arrays. Because the same biological entity is referred to by many different IDs, one needs to first map these IDs from one database to another and then query each database with its own specific IDs. This apparently trivial problem has become a challenge because various databases contain redundant information about the same biological entity. For instance, the GO categories known to be associated to a specific gene are stored in many databases such as UniProt (Bairoch *et al.*, 2005)[1], Entrez Gene (Maglott *et al.*, 2005), NetAffx (Liu *et al.*, 2003) and GO itself (Ashburner *et al.*, 2000, 2001). In spite of everybody's best efforts, because these databases are managed separately and they have different release and maintenance cycles, any data stored in more than one database creates very serious consistency and coherency problems.

A brief example will hopefully illustrate the gravity of the issues involved. Let us consider for instance, the example of a microarray experiment involving Affymetrix's GeneChips. Let us assume that a specific probe ID, 39755_at, corresponding to the human gene *XBP1*, is found to be differentially expressed. The researcher may be interested in finding the corresponding UniGene (Schuler, 1997) cluster ID for the selected probe ID, 39755_at. This can be achieved by querying NetAffx (Liu *et al.*, 2003) with the given probe ID, 39755_at, which yields the Hs.437638 cluster ID. Alternatively, one can find the cluster ID by querying NCBI's UniGene database with the gene name, *XBP1*. In this example, there exist at least two paths which yield the required information and following both paths yields the same final result. However, let us now assume that one is interested in the GO annotations associated with this gene.

Querying each of the resources above with the IDs representing the same gene, *XBP1*, yields very different results. UniProt queried with P17861 provides two unique GO terms: transcription factor activity and immune response; QuickGO queried with the same P17861 provides eight unique GO terms: protein dimerization activity, sequence-specific DNA binding, immune response, DNA-dependent regulation of transcription, transcription, DNA binding, transcription factoractivity and nucleus; NCBI's Entrez Gene entry *XBP1* provides seven unique GO terms: immune response, protein dimerization activity, sequence-specific DNA binding, DNA-dependent regulation of transcription, transcription, transcription factor activity and nucleus; PIR's iProClass (Wu *et al.*, 2003) entry P17861 provides five unique GO terms: immune response, nucleus, DNA-dependent regulation of transcription, transcription factor activity and DNA binding, whereas GO (XBP1_HUMAN) provides only two unique GO terms: immune response and transcription factor activity. Essentially, querying five different resources can yield anything between two and eight GO terms for the same gene. This situation is nothing short of disastrous. When one retrieves annotations for a set of genes from a particular source, one is always left to wonder whether the results obtained are really the entire picture or just a part of it, and whether one should continue to query other sources or just use the data retrieved so far.

Until the various resources currently available are organized into a real semantic web, free of coherency and consistency problems, arguably the best approach to retrieving annotations for a set of given biological entities is to query the authoritative source of such annotations for the given entity. In turn, in order to do this, one must map various types of IDs onto each other. This is also a tremendous challenge since various IDs can be mapped onto each other by traversing a number of alternative paths from one database to another. Since no unified map of the various databases exists, one is forced to rely on one's inherently limited personal understanding of the relationships between such databases in order to determine such a path on a case by case basis. Unfortunately, owing to the lack of global consistency and coherency, the path used to travel from one resource to another often influences dramatically the results obtained.

Another important problem is related to the cross referencing between various annotation databases. Databases such as Entrez Gene and HGNC provide gene information and are supposed to cross-reference each other. For example, gene SMCR (Smith-Magenis syndrome chromosome region) has the identifier 11113 in HGNC. The same gene is identified by Entrez Gene as gene 6600. Entrez Gene cross-references HGNC, i.e. the entry 6600 contains a field with the HGNC ID 11113. However, the reverse is not true. HGNC entry for this gene does not contain the appropriate Entrez Gene ID. Here the data are mapped only one way, from Entrez Gene to HGNC number. If the user queries HGNC using its IDs, (s)he will not be able to link to NCBI and thus will not have access to all the annotations regarding this gene available in Entrez Gene.

This problem is more widespread than one would like to believe. For instance, both UniGene and Entrez Gene focus on non-redundant genes. However, only 69.53% of the genes in UniGene can be mapped on Entrez Gene entries. Furthermore, only 43.54% of the IDs in Entrez Gene can be mapped back to UniGene. An even more striking example is the mapping between GenBank dbEST and GenPept. GenPept is supposed to contain the protein translations of the sequences in GenBank dbEST, so going back and forth between

---

[1]Swiss-Prot, TrEMBL and PIR have been recently merged as a single database in UniProt.
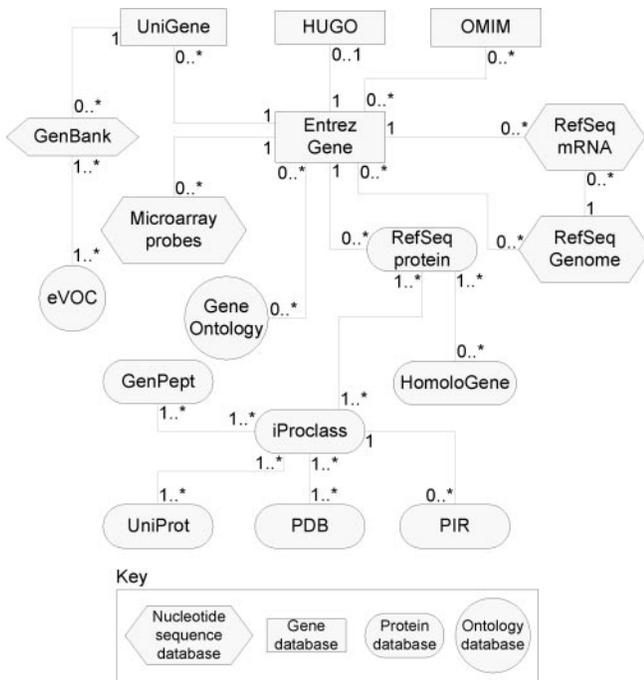
Fig. 1. OT relational database schema. This schema contains an entity for each of the source databases used by OT. The shapes represent the type of the given biological entity. A relationship between two databases is represented by a line connecting the two entities. The type of relationship between two entities is indicated by labels on the corresponding line. For instance, the relationship between Entrez Gene and Gene Ontology is many-to-many. In other words, a gene may be annotated using zero or more GO terms and a GO term may be used to annotate zero or more genes.

these resources should be trivially simple. However, this is far from being the case. At the moment, 91.6% of the entries in dbEST can be mapped to GenPept entries but the reverse mapping is possible only for 1.82% of the entries. Clearly, translations and mappings that are theoretically both meaningful and useful, cannot always be performed just by querying the resources which are supposed to allow them. These examples strongly support the idea that ID mappings cannot be done casually, by ad hoc, need-driven queries, or quick-and-dirty Perl scripts, as most researchers currently do. These quick solutions might satisfy an immediate need for *a* translation but offer no guarantees that the translation performed is the best possible mapping, nor that the results are correct or complete. At this time, the issues of incoherent name spaces between various databases represent a serious impediment to using the existing annotations at their full potential. Navigating between various such name spaces by mapping IDs from one database to another is a very important issue that must be addressed in a thorough and systematic way.

3 METHODS

In order to address the above problems, we undertook a thorough study of the following 17 annotations databases and their respective types of IDs: Swiss-Prot (IDs, accession numbers), TrEMBL (accession numbers, TrEMBL IDs), NREF (protein IDs), PIR (accession IDs), GO (GO IDs),

Table 2. Authoritative database sources in OT for different types of biological entities

Biological entity	Authoritative source
Gene	Entrez Gene
Protein	UniProt
Protein structure	PDB
Nucleotide sequence	GenBank
Nucleotide sequence cluster	UniGene
Pathway	KEGG
Disease	OMIM

A path is considered biologically valid if it either starts or ends in one of the authoritative sources. For instance, when converting from or to a gene ID, the path must start or end in Entrez Gene, respectively.

Table 3. A comparison of the scopes of OT, SOURCE, MatchMiner, GeneMerge and RESOURCERER: types of input IDs supported and number of possible translation types

Software name	Types of input IDs	Number of translations
Onto-Translate	22	462
MatchMiner	11	137
SOURCE	6	96
GeneMerge	12	30
RESOURCERER	1	16

KEGG (pathway IDs), Entrez Gene (Gene ID, gene symbol), GenBank (GI ID, accession and sequence numbers), GenPept (accession numbers), IMAGE (clone ID), RefSeq (protein, genome, mRNA accession number), UniGene (cluster ID), OMIM (OMIM number), PDB (PDB ID), Eukaryotic Promoter Database (accession number), HUGO Gene Nomenclature Committee (HGNC ID) and NetAffx (Affymetrix probe IDs). Based on the structure of these databases, we developed a relational database that allows meaningful mappings of various types of IDs onto each other. This meta-database was implemented in Oracle and all relevant data from the above databases were downloaded and used to populate the local database. Figure 1 shows a simplified schema of the part of the Onto-Tools database that is used by Onto-Translate (the complete schema includes over 70 tables). As shown in the figure, Entrez Gene, RefSeq and iProclass databases are used as central hubs that link all other source databases.

Using this database as a back-end resource, we developed a tool, Onto-Translate (OT), that can perform arbitrary translations in an optimal manner. Given two types of IDs, a translation source ID and a translation destination ID, as well as a list of specific source IDs, the algorithm calculates an optimal route between the source type and the translation type and performs the translation. The optimality of the translation is not intended in the sense of finding the translation that involves the shortest path (i.e. the lowest number of intermediate translations) but rather by the trustworthiness of the data contained in various databases. A path is defined as trustworthy if for every ID type used in any of the necessary intermediate translations, the path passes through the tables corresponding to the database that is considered as the authoritative source for that particular type of ID. For instance, Entrez Gene is considered as the authoritative source for gene data, KEGG is considered the authoritative source for pathway data, PDB the authoritative source for protein structures, etc. Thus, even if the entries of many databases across the world contain protein structure IDs, for instance, a translation involving this type of ID must use data from PDB in order to be valid.
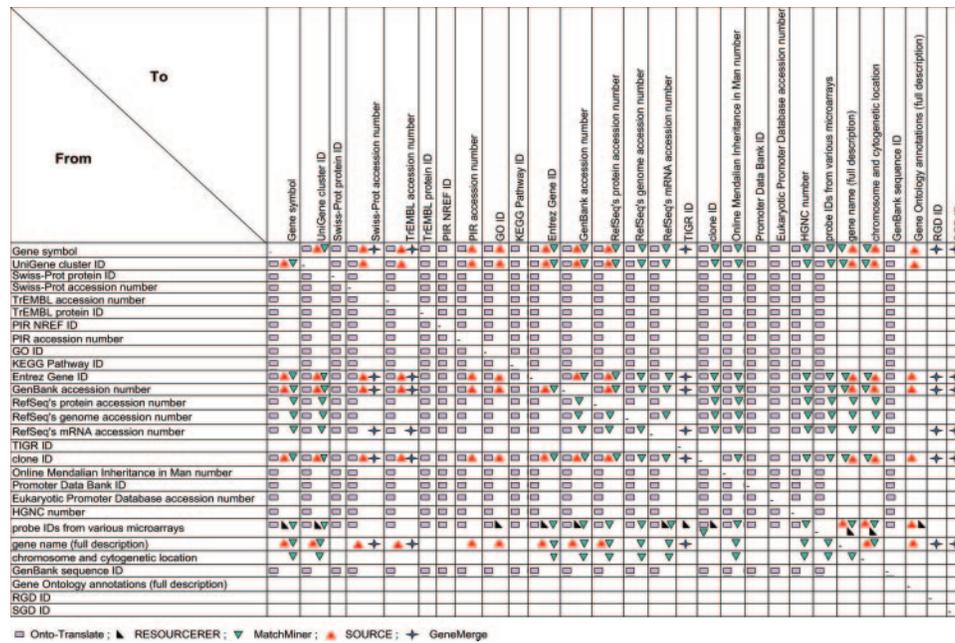
Fig. 2. A comparison of the scopes of OT, RESOURCERER, MatchMiner, SOURCE and GeneMerge. in terms of possible mappings between various types of IDs.

Table 2 shows some examples of some ID types and their authoritative sources.

The tables in the OT database and their relationships are represented as nodes and edges, respectively, in a graph structure. The basic relationships between IDs remain as given in the source databases. In a first step, the algorithm traverses the graph to find all possible paths between the source type and the destination type. This is done based on the semantic relationships between various source databases which are captured by the constraints of the Oracle database dictionary. After obtaining all possible paths between the source type and the destination type, OT removes the paths that are not trustworthy according to the criterion defined above. If the algorithm cannot find a trustworthy path between the source and the destination type, an error message is generated. If several trustworthy paths are found between the source and the destination ID, several criteria are used in order to rank them: (1) a manually curated database will always be preferred to a database containing unverified data; (2) a database containing more entries will be preferred to a database with fewer entries and (3) everything else being the same, a shorter path (involving fewer intermediate translation) will be preferred to a longer one. These criteria are also centered around biological motivations. A manually curated database reflects our preference towards accuracy rather than coverage: fewer but accurate translations are deemed preferable to a larger number of translated IDs but potentially including some incorrect translations. The second criterion above is motivated by the fact that the *a priori* probability of finding a mapping for a given ID is directly proportional to the number of entries in a database. Thus, intermediate translation through a large database is more likely to successfully find mappings for all IDs required, compared with a smaller database that might contain the same types of IDs but fewer entries. Finally, the third ranking criterion is based on the assumption that the probability of losing some IDs in each intermediate translation is non-zero and constant. In these circumstances, a shorter translation path will minimize the number of IDs lost in translation and will be better than a longer one.

Once all trustworthy paths are ranked according to these criteria, the top path between the source and the destination type is chosen as the optimal one for the required translation. At this point, OT dynamically creates a database query that follows this translation path. Besides providing an output list with
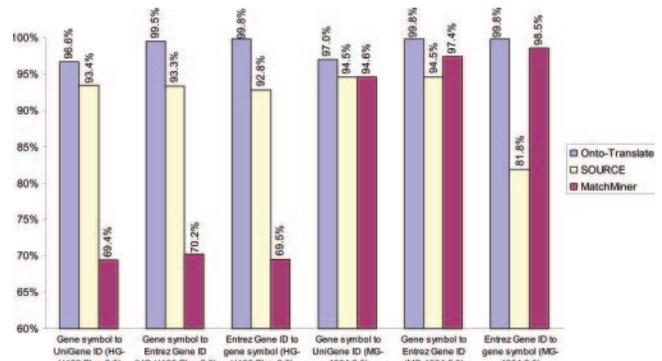
Fig. 3. A comparison of the accuracy of OT, MatchMiner and SOURCE. The input file included 19 248 gene symbols (19 562 Entrez Gene IDs) for human, and 12 991 gene symbols (13 023 Entrez Gene IDs) for mouse, from the respective Affymetrix arrays. The graph shows the percentages of the input genes successfully translated in each case.

the translation of the input IDs into the desired type of IDs, the algorithm also identifies the specific IDs which could not be translated, as well as the exact source database which broke the intermediate chain of translations required for each such specific ID. This gives the user the ability to verify that indeed the translation of that specific ID failed because the source database lacks the necessary information rather than because of a bug or missing information in our database.

Since the name-space issues that motivated the creation of OT in the first place are caused by the existence of several databases that maintain arbitrary cross-links and contain redundant information, one might ask whether the addition of yet another database would not exacerbate the problem by adding yet another level of redundancy and many more cross-references (in essence, we created cross-references from our Onto-Translate database to each of the 17 databases above). This is not the case. There are two major aspects that
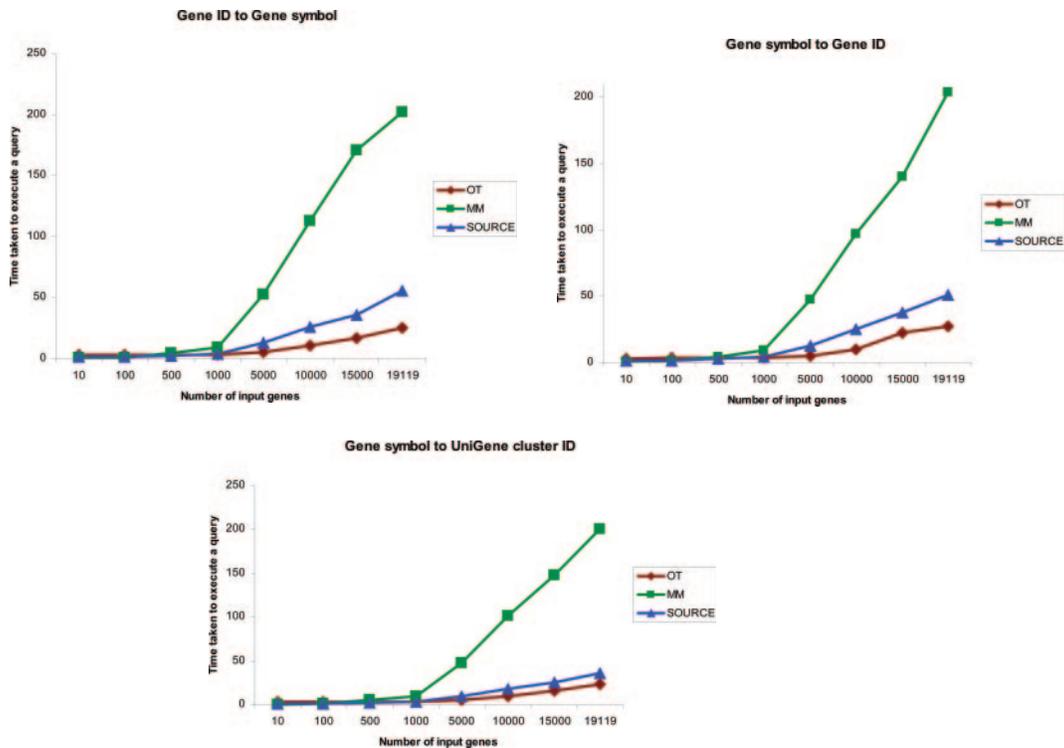
Fig. 4. Scaling properties of Onto-Translate (OT), MatchMiner (MM) and SOURCE. The graph shows the time (in seconds) necessary to translate various sets containing between 10 and 19 119 distinct genes from Affymetrix 133 Plus 2.0. At fewer than 1000 genes, the three resources have very comparable query times of <10 s. When larger sets are involved, there is a substantial performance difference.

differ between our resource and any other major resource currently available. First, most other databases are focused on either some type of biological entity (e.g. Entrez Gene for genes, UniProt for proteins, etc.) or to some specific organism (e.g. MGD for mouse, RGD for rat, etc.). In contrast, our focus is on maintaining the ID mappings themselves rather than any specific annotations. The second aspect follows from this. If a database stores annotations, the maintenance and release cycle are dictated by the evolution of the annotation activities in that area. Since the OT database does not store annotations as such, we only need to maintain the synchronization between IDs which can be done much more frequently and much more rapidly. In practice, this must be done every time any of the 17 mapped databases has a new release. In the future, this can be upgraded to an automatic overnight push of any new IDs from these databases to ours.

OT currently supports biological categories such as genes, proteins, promoters, pathways, RNAs, OMIM, ESTs and functional annotations. It can map between 29 different types of IDs which include Swiss-Prot protein ID, Swiss-Prot accession number, TrEMBL accession number, TrEMBL ID, non-redundant reference (NREF) protein ID from PIR, PIR accession ID from PIR, Gene Ontology (GO) ID, KEGG (Kanehisa *et al.*, 2002) pathway ID, GenBank GI ID, dbEST nucleotide accession number, Entrez Gene ID, Gene symbol, GenBank (Benson *et al.*, 2005) dbEST (Boguski *et al.*, 1993) nucleotide accession number, GenPept protein accession number, RefSeq's protein, genome, mRNA accession number, UniGene cluster ID, clone IDs from UniGene, OMIM number, allelic variant from OMIM, Protein Data Bank (PDB) ID, Eukaryotic Promoter Database (EPD) accession number, EPD ID, HGNC ID and probe IDs from commercial microarrays such as Affymetrix arrays, Agilent Technologies arrays, Amersham's CodeLink arrays, SuperArray, etc. The Onto-Translate tool is implemented in Java as a web application, fully integrated with the Onto-Tools (Drăghici *et al.*, 2003a,b,c; Khatri *et al.*, 2004, 2002, 2005; Khatri and Drăghici, 2005).

4 RESULTS AND DISCUSSION

Clearly, the need for a reliable way of mapping IDs from one database to another has been felt in the past. In response to this needs, several approaches have been proposed to deal with this issue although none of them addressed the problem to its full extent. The best known resources that are currently able to perform a non-trivial mapping of various biological entities are SOURCE (Diehn *et al.*, 2003) from Stanford University, MatchMiner (Bussey *et al.*, 2003) from NCI, RESOURCERER (Tsai *et al.*, 2001) from TIGR, and GeneMerge (Castillo-Davis and Hartl, 2002) from Harvard.

We compared Onto-Translate with each of these existing resources in terms of scope, accuracy of translation, speed and scaling capabilities. We define scope as the number of different mappings between types of IDs supported by a given resource. The comparison in Table 3 shows that OT has vastly larger capabilities compared with any of the existing resources. Figure 2 shows the specific translations that can be performed by each of the resources considered. Again, the difference in scope is striking.

Of course, the scope is irrelevant if the accuracy of the mappings performed is inadequate. In order to compare the accuracy of the existing resources, we performed a number of translations using OT, SOURCE, and MatchMiner (top 3 in terms of scope), and compared the number of input IDs correctly mapped by each resource for each dataset. OT consistently mapped more correct input IDs than both SOURCE and MatchMiner. The sets of genes to be translated were taken from popular human and mouse Affymetrix arrays. The set of genes contained on the HG-U133 Plus 2.0 array was used to test the

translations from gene symbols to UniGene IDs, gene symbols to Entrez Gene IDs and Entrez Gene IDs to gene symbols. Finally, for the translations involving mouse genes, we used the set of genes contained on the MG-430A 2.0 arrays. These genes were translated from gene symbols to UniGene IDs, gene symbols to Entrez Gene IDs and Entrez Gene IDs to gene symbols. Figure 3 shows a comparison of the accuracy of these translations. OT was the most accurate resource in all cases, with accuracies between 96 and 99%. For human data, SOURCE is second best with an accuracy hovering around 93%. MatchMiner is weaker with an accuracy of ∼70%. For mouse data, MatchMiner is better than SOURCE: 94–98% for MM, compared with 81–94% for SOURCE.

Figure 4 shows a comparison of the time (in seconds) necessary to perform a sample translation from gene symbols to gene IDs with Onto-Translate, MatchMiner and SOURCE. The time necessary to translate fewer than 1000 genes is approximately the same for the three resources. However, when longer lists are involved, OT is ∼2 times faster than SOURCE and ∼10 times faster than MatchMiner, in all translations performed.

5 CONCLUSIONS

This paper discusses various issues related to name space inconsistencies between existing annotation databases. The distribution of our knowledge over several databases forces researchers to navigate from one such database to another, in order to construct the correct interpretation of any given experiment. Currently, the lack of the ability to map correctly various IDs from one DB to another creates very substantial problems in annotation retrieval. We have developed a resource that addresses this stringent need. This resource includes a back-end database as well as a web tool, Onto-Translate, that provides a convenient user interface. Currently, OT is able to perform 462 types of mappings between 29 different types of IDs from 17 databases concerning 53 organisms. This is better than the other resources we have investigated in terms of: (1) number of translations possible, (2) types of IDs supported, (3) accuracy and (4) speed. OT is a part of Onto-Tools, which is freely available at http://vortex.cs.wayne. edu/Projects.html.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Ashburner,M. *et al.* (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.

Bairoch,A. *et al.* (2005) The universal protein resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Benson,D.A. *et al.* (2005) Genbank. *Nucleic Acids Res.*, **33**, D34–D38.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Boguski,M.S. *et al.* (1993) dbEST—database for expressed sequence tags. *Nat. Genet.*, **4**, 332–333.

Bussey,K.J. *et al.* (2003) Matchminer: a tool for batch navigation among gene and gene product identifiers. *Genome Biol.*, **4**, R27.

Castillo-Davis,C.I. and Hartl,D.L. (2002) GeneMerge—post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, **19**, 891–892.

Diehn,M. *et al.* (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31**, 219–223.

Drǎghici,S. *et al.* (2003a) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, **31**, 3775–3781.

Drǎghici,S. *et al.* (2003b) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Drǎghici,S. *et al.* (2003c) Assessing the functional bias of commercial microarrays using the Onto-Compare database. *BioTechniques*, 55–61.

Kanehisa,M. *et al.* (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.

Khatri,P. and Drǎghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Khatri,P. *et al.* (2002) Profiling gene expression using Onto-Express. *Genomics*, **79**, 266–270.

Khatri,P. *et al.* (2004) Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res.*, **32**, W449–W456.

Khatri,P. *et al.* (2005) Recent additions and improvements to the Onto-Tools. *Nucleic Acids Res.*, **33**, W762–W765.

Liu,G. *et al.* (2003) Netaffx: affymetrix probesets and annotations. *Nucleic Acids Res.*, **31**, 82–86.

Maglott,D. *et al.* (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **30**, 137–140.

Schuler,G.D. (1997) Pieces of puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.

Tsai,J. *et al.* (2001) Resourcerer: a database for annotating and linking microarray resources within and across species. *Genome Biol.*, **2**, software0002.1–software0002.4.

Wu,C.H. *et al.* (2003) The protein information resource. *Nucleic Acids Res.*, **31**, 345–347.