Standardized Representations of the Literature: Combining Diverse Sources of Ribosomal Data

Russ B. Altman, Neil F. Abernethy & Richard O. Chen
Stanford Section on Medical Informatics
SUMC, MSOB X-215, Stanford, CA, USA, 94305-5479
(415) 725-3394, fax: (415) 725-7944, {rba, nfa, rchen}@smi.stanford.edu

From: ISMB-97 Proceedings. Copyright © 1997, AAAI (www.aaai.org). All rights reserved.

Abstract

We are building a knowledge base (KB) of published structural data on the 30s ribosomal subunit in prokaryotes. Our KB is distinguished by a standardized representation of biological experiments and their results, in a reusable format. It can be accessed by computer programs that exploit the rich interconnections within the data. The KB is designed to support the construction of 3D models of the 30S subunit, as well as the analysis and extension of relevant functional and phylogenetic information. Most published information about the structure of the ubiquitous ribosome focuses on *E. coli* as a model system. At the same time, thousands of RNA sequences for the ribosome have been gathered and cataloged. The volume and complexity of these data can complicate attempts to separate structural data peculiar to *E. coli* from data of universal relevance. We have written an application that dynamically queries the KB and the Ribosome Database Project, a repository of ribosomal RNA sequences from other organisms, in order to assess the relevance of structural data to particular organisms. The application uses the RDP alignment to determine whether a set of data refer primarily to conserved, mismatched , or gapped positions. For a set of 16 representative articles evaluated over 211 sequences, 73% of observations have unambiguous translations from *E. coli* to the other organisms, 21% have somewhat ambiguous translations, and 6% have no translations. There is a wide variation in these numbers over different articles and organisms, confirming that some articles report structural information specific to *E. coli* while others report information that is quite general.

Introduction

The effect of the world wide web (WWW) on the dissemination of biological information has been remarkable. Access to databases containing nucleic acid sequences, protein sequences, protein structure, information about metabolism, genetic information, and other specialized areas has become nearly universal, and now provides a reliable way to publish and retrieve information (see, for example http:/molbio.info.nih.gov/molbio/db.html). In addition to the growth of online databases, there has also been a growth in the availability of published scientific reports, in the form of online journals [1], online conference proceedings [2], and even "self-publishing" of work by individual investigators. In the case of peer reviewed work of relevance to biomedicine, there are now powerful bibliographic databases that contain basic information about most published articles—including some with full text [3].

The volume of published literature is increasing so rapidly, however, that it is difficult for individuals to track important developments within their field, and nearly impossible for them to stay abreast of important developments in more distant fields. It is particularly difficult to effectively search the published prose literature. Searching with proxies for content such as author, keywords or other standard bibliographic information can be successful, but can also have significant rates of false positive retrievals. Beyond searching, investigators may soon want to analyze the literature with computational tools that assist in the filtering and analysis of published information. Current capabilities in natural language understanding are generally not sufficient to allow computer programs to understand the key content of scientific articles, and use this information to assist investigators in the formulation and testing of scientific hypotheses[3].

For these reasons, there has been interest in developing formal methods for representing, storing and communicating biological knowledge[4]. If the standard bibliographic information can be supplemented with computer-readable representations of the experimental system, the protocols used, and the results, then it becomes feasible to create a wide array of applications that use this information. Such applications could support more sensitive literature searching, automatic consistency checking and even automatic hypothesis testing. In some ways, the current sequence and structure databases serve as successful examples of formal methods for communicating biological knowledge. In these cases, the details of particular experimental techniques have been standardized so that data files can represent the key elements of the experiment (GenBank records contain the results of sequencing experiments, Protein Data Bank records contain the results of xray crystallographic or NMR experiments). Many resources are now taking advantage of these standardized formats to link the data from multiple sources [5, 6]. However, the vast majority of scientific observations reported in the literature contain findings that do not fit naturally into one of these databases. There are

often not enough observations of any single type to justify an entire database, and so these findings are typically not available in an accessible, computer-readable format.

We are testing the viability of structured representations of these other parts of the biomedical literature, and have identified a narrow domain of investigation for which a knowledge base of the pertinent literature would be useful. The problem of determining the structure of the ribosome (the site of mRNA translation into protein) is a good testbed [7]: it is of critical biological importance, has a substantial and nontrivial published literature, has numerous active research groups, and there are significant outstanding questions and disagreements about the current scientific models and their adequacy. The 30S subunit is a large molecule made of RNA and protein components. It is the site of mRNA translation into protein, and it has been the subject of numerous structural and functional studies over the past 30 years. More specifically, there have been six structural models of the ribosome published in the last 15 years, all of which have some significant similarities and differences [8-13]. The union of the bibliographies of the papers reporting these models consists of roughly 240 peer-reviewed articles. These include numerous reports of biochemical, biophysical and genetic studies that shed light on the structure.

In this paper, we make an initial report of a KB containing structured representations of the 30s ribosome literature. We describe the format of the KB, and our methodology for entering information. The KB contains information primarily about the 30S ribosomal subunit in *E. coli* (the dominant experimental system in this field). In order to demonstrate utility, we have written an application that evaluates the relevance of any article in the KB to the task of building a model of the 30S subunit for a particular organism. This task is difficult because it requires information about what the article reports about *E. coli* (contained in our KB) as well as how to relate the *E. coli* sequence to a different organism (contained in the Ribosomal Database Project—RDP—resource [14]). As such, our results demonstrate the promise of explicit declarative representations of the literature—especially in combination with other online information resources.

Methods

The methods we employ can be divided into three areas: constructing the KB, integrating our application with the KB and the RDP resources, and the performing an analysis of the relevance of KB articles to multiple organisms.

Building the Knowledge Base

The KB is implemented in the ONTOLINGUA frame-based representation language [15]. ONTOLINGUA was selected because it offers 1) a frame-based representation system in which a taxonomy of classes can be elaborated, specifying necessary attributes (and the range of possible values for each attribute) for each class, including links between classes, 2) a WWW interface which allows editing and browsing concepts from anywhere on the internet, 3) programmatic access for reading and writings its contents, and 4) a group of developers who are interested in understanding how large KBs should be supported. At the leaves of the class hierarchy are the instances of each concept, which represent the actual facts reported in the literature. There are many other possible choices for the KB framework, and the choice is somewhat arbitrary at this prototyping stage [16].

The ONTOLINGUA KB always has the object `Thing` at the root. We define five major subclasses of concepts for our KB. The first three subclasses organize how we think about the physical objects involved in the domain, and the next two organize how we think about data and publications.

1. We define a concept `molecule`, which includes all molecules that can exist as separate entities in a solution. This includes subclasses, for example, of `nucleic acids`, `proteins`, `ions`, and `small-organic` molecules. In general, molecules are defined by covalent links. Molecules have attributes such as `name`, `weight`, and `molecular-components` (described below). Particular classes have more detailed information (such as an attribute `length` for proteins and nucleic acids).

2. We define a concept `molecular-ensemble`, that is defined as one or more molecules associated non-covalently, such as a complex of RNA and proteins or a protein with its small organic ligand. The 30S ribosomal subunit, therefore, is an ensemble consisting of an RNA molecule and 21 proteins. Ensembles must have (at a minimum) a name and a list of parts (a set of molecules).

3. We define a concept `molecular-component`. Unlike molecules and molecular ensembles, which are strictly defined, molecular components are set of terms that are used to label parts of a `molecule` or a `molecular-ensemble`, but are not defined as separate physical entities. Molecular-components include the amino acid components of a protein polypeptide (which are, after all, no longer amino acids when they occur within the polypeptide!), the RNA bases of an RNA chain, the named binding sites of proteins, and other definable molecular labels to which one might refer.

4. We define a concept `data`. Data are individual pieces of information that can be reported in the literature. The subclasses are organized into the types of data found in the ribosomal structural literature, including `biochemical-data`, `biophysical-data`, `genetic-data` and `phylogenetic-data`. Thus, for example, chemical-cross-linking and enzymatic-

footprinting experiments produce `biochemical-data`, NMR and neutron-diffraction experiments produce `biophysical-data`, and analyses of covariation between columns of a multiple alignment produces `phylogenetic-data`. These distinctions are clearly subject to opinion, and must, in general, be exposed to peer review. As a rule, each subclass within the data taxonomy is associated with a particular type of experiment, and has the attributes necessary to specify the major findings of that type of experiment. Many of these attribute values are instances of `molecules`, `molecular-ensembles`, or `molecular-components`. Others may be strings or numeric constants. Thus, for example, a chemical-crosslinking experiment has a `target-complex`, a list of the molecular-components present in the experiment, the `crosslinking-agent`, used to link the components, and the `cross-linked-objects` of the complex that are cross-linked.

5. We define a concept `reference-information` that has subconcepts such as `authors`, `institutions` and `publications` relevant to the 30S ribosomal subunit. One of the most significant subclasses of `publications` is `peer-reviewed-journal-article`. In addition to all the expected bibliographic attributes of a journal article, we have a critical attribute called `reports` that is filled by a list of the data instances that are reported by the article in question. These data instances have, in turn, an inverse attribute called `reported-by` that points back at the article. This is the key link in the system, because it establishes the relationship between a bibliographic entry, and a structured representation of the data reported by this article. By following this link, and looking at the details of the data instances, we have full knowledge of the objects upon which measurements were made, and what the measurement results were.

Given a new article, we determine the sections of the KB concept taxonomy to which the reported data belongs. If we are unable to find such data, then we add new classes in the appropriate part of the tree, and the new class is assigned a set of required attributes (e.g., cross-linking experiments always must have a `cross-linking-agent`). We create one instance of a class for every observation reported in the paper. This is often found in the Results section or in figures and tables of the paper. We always attempt to represent the paper as originally reported. That is, we make no assumptions about the quality of the experiments or data, leaving that for peer-review and individual interpretation. Like the literature, the KB can contain inconsistent data, but its structure allows algorithmic approaches for the detection of these inconsistencies.

Integrating the application with KB and RDP

In order to test the utility of our current implementation, we built an application with the goal of evaluating the relevance of a article in the KB to any particular organism in RDP—an online collection of ribosomal sequence information, including alignments of the RNA components of the 30S subunit. The structure of the 30S subunit has been studied chiefly in the context of *E. coli* as a model system, and the 30S subunits from other organisms are known to share significant structural similarities [17]. Yet there are differences between the sequences of all these organisms, and it is not always clear if the measurements reported in a study on *E. coli* are relevant to other organisms. Because the RDP has good quality alignments between all the RNA sequences of the 30S subunit, and because our KB has a detailed representation for each article of what measurement was made and what components of the 30S were involved, we have an opportunity to combine these two resources to measure the potential relevance of an article to a new organism.

Our prototype application was built in Visual Basic for Windows NT. It uses a plug-in WWW browser to provide the user with the pre-existing RDP interface. Unlike a normal browser, however, the application filters incoming text in the GenBank file format. This allows the user to fetch information from RDP about available organisms and their sequences. The user also has the options of loading sequences from a file, pasting them from another source, or accessing them from a local database. To link these external data sources with the KB, the application communicates with ONTOLINGUA via an http-based interface that allows querying over the network. Results are displayed graphically and written to a local file for further manipulation by the user. The user selects an organism and a set of KB articles, and the application summarizes the relevance of the article to the organism or set of organisms available.

Assessing Relevance of an Article to an Organism

The relevance of an article to an organism depends on the following data: (1) The alignment of the organism RNA sequence with the sequence of the organism actually used in the experiment (usually *E. coli*, although not limited to this). This was retrieved from RDP at http://rdp.life.uiuc.edu/, and (2) The list of data items reported by the article of interest. Given the article, we traced the `reports` link to one or more basic data items within the KB. For each of these items, we sought attributes whose values were individual RNA bases in *E. coli* or short runs of bases within the sequence. The application makes a sequence of calls to the KB to obtain a list of all RNA bases for which information is reported in the articles.

With these two pieces of information, the application computes three types of information for each article. First, it determines which bases referenced in the article have exact sequential identities in the alignment with E. Coli. It labels these base relations type IA—they are most likely to play similar structural roles in the new organism. Second,

it determines those bases referenced in the article that do not have exact identities in the multiple alignment (mismatches). It labels these base relations type IB—they are somewhat less likely to have the same structural role in the new organism, but can still be used to translate information from *E. coli* into the new organism. Taken together, type IA and IB relations constitute the set of all data for which any possible interpretation exists for the second sequence. Finally, the application determines which bases in the article correspond to deletions or otherwise unaligned segments of the new sequence, and thus have no structural interpretation in the new organism, and label these type II.

For these experiments, we selected 16 articles from the KB which represent information of structural relevance to the 30S subunit [18-32]. We took 211 16S RNA sequences for different organisms (selected from each of the phylogenetic subheadings as organized within RDP), and computed the total number of bases referenced in each article, and the breakdown of these bases as type IA, IB and II matches with the new organism. In order to assess the significance of type IA, IB and II matches in the reported data, we computed as a baseline the percent identity, mismatch, and deletions for all positions and organisms in the complete sequence alignment.

Results

Currently, the KB contains a concept hierarchy with roughly 120 classes, and a total of 5000 instances (including molecular components, articles, and representations of the data). There are about 100 articles represented partially, of which the contents of 25 are fully represented. Figure 1 shows a part of the class hierarchy of our knowledge base, focusing on biochemical data. Figure 2 shows a typical instance of the class of `journal-articles`, and an instance of the data it reports. Figure 3 shows the secondary structure of the 16S ribosome annotated with information from RDP to show differences in the sequence between *E. coli* and *Msr. thmoph*, and the areas of relevance to one of the articles.

Table 1 shows the 16 articles that we analyzed , the number of specific base references contained in the data instances that are reported by each article, and the breakdown of these bases with respect to their alignment with *E. coli*. Table 2 shows, for a particular article, the breakdown over a representative set of species. Table 3 shows, for each article tested, the organisms with the most mismatches and deletion—indicating that these articles are least relevant to these species. The full data set is at http://www-smi.stanford.edu/projects/helix/pubs/ismb97-aac/.

Discussion

One of the advantages of working on model systems in biology is the ability to focus effort on a single experimental system, understand its workings, and then generalize to other systems. Our experiment is an example of one such generalization. We have estimated that, on average, roughly 94% of the structural data reported for *E. coli* can be translated into meaningful constraints for other organisms, either as type IA or IB constraints. Of course, the strict validity of this estimate would require that the laboratory experiments be repeated with the different organisms, in order to confirm the observations. Unfortunately, the time and expense of such an undertaking is prohibitive. However, investigators with a specific interest in other organisms can get a preliminary indication of the quantity of relevant data available for modeling these related molecules, and what parts of the *E. coli* data set will not be useful. The idea of using an *E. coli*-specific knowledge base to generate a data set "by analogy" is not new, and has been used to estimate the metabolic pathways of H. Influenza [33]. We are developing a system to compute structural models by translating the data in our KB into structural constraints suitable for algorithms that compute three-dimensional structure. An extension of the application reported here will allow us to generate data sets for related organisms.

For this experiment, we used a simple approach to evaluating the relevance of an article. First, we assumed that the relevance of an article to a new organism is based only on degree of conservation at individual base positions, as measured by the similarity of the bases at the positions aligned to *E. coli*. In order to justify this assumption, we chose a set of articles that report the proximities of bases to one another or to other elements of the 30S subunit (such as the S-proteins). Thus, for these articles, using a list of referenced bases is a reasonable measure of the substructures studied. Second, we summarized the relationship of an article to an organism by simply counting the fraction of referenced bases in *E. coli* that are exact matches, mismatches, or the site of deletions in the other organism. We reasoned that measurements involving strictly conserved bases are likely to be good candidates for translation into new organisms—especially given the relatively strict conservation of secondary structure. On the other hand, measurements involving bases that are aligned but mismatched may also be reliably translated, but may have a greater chance of being less relevant to the new organism. Finally, measurements involving bases that are deleted in *E. coli* are of little relevance to the other organism, and probably reflect a feature particular to *E. coli*.

On average, the *E. coli* 16S RNA shares 71% identity with the other 16S RNA sequences used in this study, with 21% aligned mismatches, and 8% gaps. If the published data

were gathered on segments of RNA randomly drawn from the *E. coli* sequence, then this breakdown would be expected for our Type IA, IB, and II statistics for the articles that were analyzed. However, our results show that articles reporting information about smaller subsets of bases tend to report a higher percentage of Type IA constraints. These articles focus on structurally significant interactions that tend to involve conserved bases.

In building our KB of the published literature on the 30S ribosome, we are forced to make modeling decisions. We collaborate in these decisions with other RNA biologists in an attempt to make reasonable choices. Nonetheless, our taxonomy of data types remains subjective, and can be criticized in its details. In order to demonstrate the feasibility and utility of representing biological information in this manner, we are creating this first prototype knowledge base. If the prototype is successful, two outstanding issues remain: Who should decide on the proper taxonomy for the biological literature? Given a taxonomy, how can information be entered into a knowledge base? We have not answered these questions, but suggest that a panel of peer reviewers can be assembled, for any subdiscipline, to ensure that the KB classes are reasonable. For data entry, it will be critical to move this task to the authors of scientific papers. We suggest that professional journals can require that manuscripts be accompanied by a set of knowledge base instances that are reviewed along with associated manuscripts. Successful review would then lead to both publication of the manuscript, and addition of the incremental knowledge base instances. In this way, a shared and computable representation of the published literature would grow with the literature itself. There has been good progress in the creation of tools that allow entry of data in standard formats[34], and techniques are being developed to acquire information for KBs similar to ours [35].

The main challenge in designing the classes within the KB is deciding the proper granularity of representation. Clearly, a scientific article contains many details of potential interest, and it is not yet feasible to represent everything in our KB. We try to make sure that our representations capture the type of experiment being performed, the key components of the experiment, and the results. We do not stress the detailed reagents, concentrations and laboratory techniques, although these have been the subject of investigation by others [36]. We have stressed physical and structural data so far, and have not addressed the issue of representing functional information, a more difficult task, and one that is also receiving attention [37].

Our policy of representing all articles (and their contents) from the perspective of their authors is critical. We do not make any judgments about the reliability or consistency of the articles. Instead, the declarative representation of the literature allows us to evaluate suspicious data in light of other evidence in the KB. In separate work, we have used our KB as a source of data for building 3D models of the 30S ribosomal subunit. We are able to evaluate our models by testing them against the data within the KB—both the data used to compute the model, and other data not used, but deemed to be relevant.

There are many other potential uses of the KB. In this report, we have shown its utility in combining alignment information with structural experimental information. Storing data in a structured format may make computational tasks such as literature searching, error checking, and model building easier by providing a reliable network of information for programs and programmers to build upon. Our application demonstrates our ability to create links to diverse programmatic and data resources, including the KB, the WWW-based RDP, Genbank, an RNA secondary structure viewer, and even a separate initial version of our KB. Although many other programs and databases exchange data in pre-defined format, often they contain implicit models about the data. The knowledge base itself provides a declarative model of information around which programs can be structured, providing an explicit context for both the input and output of applications which are built upon it.

Acknowledgments

We thank Ramon Felciano and Harry Noller for useful discussions, and James Rice for his support of ONTOLINGUA. This work is supported in part by NIH-LM05652, LM06442, NSF DBI-9600637, and a grant from IBM.

References

1. See, for example, http://www.highwire.org/
2. See, for example, http://www-smi.stanford.edu/people/altman/psb97/index.html
3. Schatz, B.R., *Information Retrieval in Digital Libraries: Bringing Search to the Net.* Science, 1997. **275**: p. 327-334.
4. Hafner, C.D., *et al.*, *Creating a knowledge base of biological research papers*, in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, R. Altman, *et al.*, Editors. 1994, AAAI Press: Menlo Park. p. 147-155.
5. Etzold, T., A. Ulyanov, and P. Argos, *SRS: Information Retrieval System for Molecular Biology Data Banks.* Methods in Enzymology, 1996. **266**: p. 114.
6. Sonnhammer, E. and R. Durbin, *A Workbench for large-scale sequence homology analysis.* CABIOS, 1994. **10**(3): p. 301-307.
7. Hardesty, B. and G. Kramer, *Structure, Function, and Genetics of Ribosomes*. 1986, New York: Springer Verlag. 87-100.

8. Fink, D.L., *et al.*, *Computational methods for defining the allowed conformational space of 16S rRNA based on chemical footprinting data.* RNA, 1996. **2**: p. 851-866.
9. Expert-Bezancon, A. and P.L. Wollenzien, *Three-dimensional arrangement of the Escherichia coli 16 S ribosomal RNA.* J Mol Biol, 1985. **184**(1): p. 53-66.
10. Malhotra, A. and S.C. Harvey, *A Quantitative model of the E. coli 16S RNA in the 30S Ribosomal Subunit.* J Mol Bio, 1994. **240**: p. 308-340.
11. Hubbard, J.M. and J.E. Hearst, *Computer modeling 16 S ribosomal RNA.* J Mol Biol, 1991. **221**(3): p. 889-907.
12. Stern, S., B. Weiser, and H.F. Noller, *Model for the three-dimensional folding of 16 S ribosomal RNA.* J Mol Biol, 1988. **204**(2): p. 447-81.
13. Brimacombe, R., *et al.*, *A detailed model of the three-dimensional structure of Escherichia coli 16 S ribosomal RNA in situ in the 30 S subunit.* J Mol Biol, 1988. **199**(1): p. 115-36.
14. Maidak, B.L., *et al.*, *The Ribosomal Database Project (RDP).* Nucleic Acids Research, 1996. **24**: p. 82-85.
15. Farquhar, A., *et al.*, *Collaborative Ontology Construction for Information Integration.*, . 1995, Knowledge Systems Laboratory Department of Computer Science.
16. Karp, P.D., *The design space of frame knowledge representation systems*, . 1992, SRI International Artificial Intelligence Center.
17. Noller, H.F. and C.R. Woese, *Secondary structure of 16S ribosomal RNA.* Science, 1981. **212**(4493): p. 403-11.
18. Moazed, D. and H.F. Noller, *Binding of tRNA to the ribosomal A and P sites protects two distinct sets of nucleotides in 16 S rRNA.* J Mol Biol, 1990. **211**(1): p. 135-45.
19. Moazed, D., S. Stern, and H.F. Noller, *Rapid chemical probing of conformation in 16 S ribosomal RNA and 30 S ribosomal subunits using primer extension.* J Mol Biol, 1986. **187**(3): p. 399-416.
20. Powers, T., *et al.*, *Probing the assembly of the 3' major domain of 16 S ribosomal RNA. Quaternary interactions involving ribosomal proteins S7, S9 and S19.* J Mol Biol, 1988. **200**(2): p. 309-19.
21. Powers, T., *et al.*, *Probing the assembly of the 3' major domain of 16 S rRNA. Interactions involving ribosomal proteins S2, S3, S10, S13 and S14.* J Mol Biol, 1988. **201**(4): p. 697-716.
22. Powers, T. and H. Noller, *Hydroxyl radical foot printing of ribosomal proteins on 16S rRNA.* RNA, 1995. **1**: p. 194-209.
23. Svensson, P., *et al.*, *Interaction of ribosomal proteins, S6, S8, S15 and S18 with the central domain of 16 S ribosomal RNA.* J Mol Biol, 1988. **200**(2): p. 301-8.
24. Osswald, M., *et al.*, *RNA-protein cross-linking in Escherichia coli 30S ribosomal subunits; determination of sites on 16S RNA that are cross-linked to proteins S3, S4, S5, S7, S8, S9, S11, S13, S19 and S21 by treatment with methyl p-azidophenyl acetimidate.* Nucleic Acids Res, 1987. **15**(8): p. 3221-40.
25. Greuer, B., *et al.*, *RNA-protein cross-linking in Escherichia coli 30S ribosomal subunits; determination of sites on 16S RNA that are cross-linked to proteins S3, S4, S7, S9, S10, S11, S17, S18 and S21 by treatment with bis-(2-chloroethyl)-methylamine.* Nucleic Acids Res, 1987. **15**(8): p. 3241-55.
26. Stiege, W., *et al.*, *Intra-RNA cross-linking in Escherichia coli 30S ribosomal subunits: selective isolation of cross-linked products by hybridization to specific cDNA fragments.* Nucleic Acids Res, 1988. **16**(10): p. 4315-29.
27. Stern, S., R.C. Wilson, and H.F. Noller, *Localization of the binding site for protein S4 on 16 S ribosomal RNA by chemical and enzymatic probing and primer extension.* J Mol Biol, 1986. **192**(1): p. 101-10.
28. Stern, S., *et al.*, *Interaction of proteins S16, S17 and S20 with 16 S ribosomal RNA.* J Mol Biol, 1988. **200**(2): p. 291-9.
29. Stern, S., *et al.*, *Interaction of ribosomal proteins S5, S6, S11, S12, S18 and S21 with 16 S rRNA.* J Mol Biol, 1988. **201**(4): p. 683-95.
30. Stern, S., *et al.*, *RNA-protein interactions in 30S ribosomal subunits: folding and function of 16S rRNA.* Science, 1989. **244**(4906): p. 783-90.
31. Zwieb, C. and R. Brimacombe, Uridine 1239 in the 16S RNA sequence is cross-linked to protein S7. Nucleic Acids Res, 1979. **6**(5): p. 1775-90.
32. Haselman, T, Camp, D.G., Fox, G.E., Phylogenetic evidence for tertiary interactions in 16S-like ribosomal RNA, Nucleic Acids Res, 1989. 17(6): P. 2215-21.
33. Karp PD; Ouzounis C; Paley S . HinCyc: a knowledge base of the complete genome and metabolic pathways of H. influenzae. Ismb, 1996, 4:116-24
34. See, for example, http://www.ncbi.nlm.nih.gov/ Sequin/ and http://terminator.pdb.bnl.gov:4148/autodep-basepage.html
35. Tu SW; Eriksson H; Gennari JH; Shahar Y; Musen MA . Ontology-based configuration of problem-solving methods and generation of knowledge-acquisition tools: application of PROTEGE-II to protocol-based decision support. Artificial Intelligence in Medicine, 1995 Jun, 7(3):257-89.
36. Baclawski, K., *et al.*, *Database Techniques for Biological Materials and Methods*, in *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, L. Hunter, D. Searls, and J. Shavlik, Editors. 1993, AAAI Press: Menlo Park. p. 21-28.
37. Karp, P.D. and Riley, M. Representations of Metabolic Knowledge. Ismb, 1993, 1:207-215.

Table 1. Summary statistics for 16 articles published about *E. coli* structure and assessed for relevance to 211 non-*E. coli* 16S RNA sequences. The journal, first author, volume and page number are used to identify each article, and the table shows the number of bases that each article provides information about (#Bases), the overall percentage of identical matches at those base positions when the 211 sequences are compared with *E. coli* (%ID), the overall percentage of mismatches (%MM), and the overall percentage of gaps in the alignments (%Gap). The density of data (#Bases) varies widely for the articles, as does the percentage of information that can be reliably translated from *E. coli* to new organisms (%ID).

Authors	Journal	Vol.	Page	# Bases	% ID=IA	%MM=IB	% Gap=II
Moazed, et. al.	JMB	187	399	1465	72%	21%	7%
Gutell, et. al.	NAR	21	3051	766	65%	27%	9%
Powers, et. al.	RNA	1	194	592	79%	18%	3%
Stiege, et. al.	NAR	16	4315	137	82%	17%	2%
Powers, et. al.	JMB	200	309	84	84%	11%	4%
Stern, et. al.	JMB	201	683	79	86%	10%	4%
Svensson, et. al.	JMB	200	301	66	78%	22%	0%
Stern, et. al.	JMB	200	291	53	86%	11%	3%
Stern, et. al.	JMB	192	101	39	80%	18%	2%
Geuer, et. al.	NAR	15	3241	39	54%	18%	28%
Powers, et. al.	JMB	201	697	37	88%	10%	2%
Stern, et. al.	Science	244	783	32	84%	16%	0%
Moazed, et. al.	JMB	211	135	18	80%	3%	17%
Osswald, et. al.	NAR	15	3221	18	89%	9%	2%
Camp, et. al.	NAR	17	2215	16	72%	13%	16%
Brimacombe, et. al.	NAR	6	1775	3	98%	1%	1%

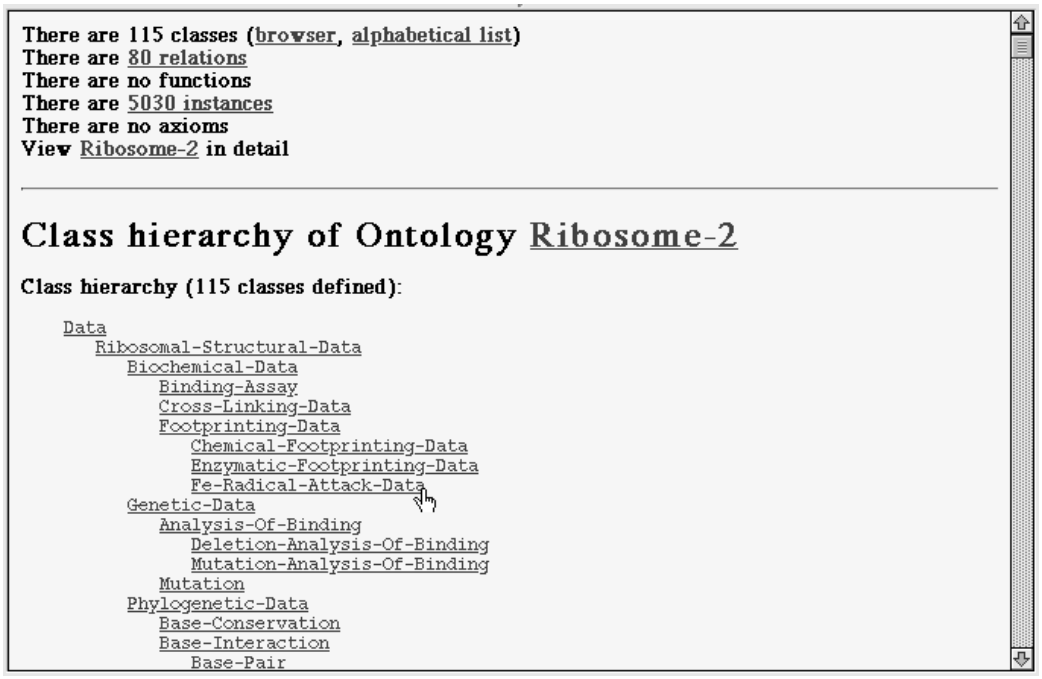
Table 2. For a particular article from Table 1 (jmb-moazed-187-399), a sample set of species are given along with the number of bases in the 16S sequence of the new species, the number of bases that align with *E. coli* 16S, and the number and percent of bases from jmb-moazed-187-399 that are identical matches, aligned mismatches and gaps. Jmb-moazed-187-399 is most relevant to Flx. polymo (total alignable bases of 1447, only 6% of sequence gapped with respect to *E. coli*), and least relevant to Mccl.10mfx (total alignable bases 1100, 29% of sequence gapped).

Organism	Sequence length	Alignable Bases	ID		MM		Gap	
Flx.polymo	1476	1447	1046	68%	401	26%	95	6%
Brv.andrsn	1450	1442	1061	69%	381	25%	100	6%
Bac.putred	1468	1440	1049	68%	391	25%	102	7%
Ric.ricket	1443	1430	1100	71%	330	21%	112	7%
Blt.sp1	1453	1420	1012	66%	408	26%	122	8%
Afp.felis	1426	1416	1112	72%	304	20%	126	8%
Zoo.ramig2	1407	1401	1106	72%	295	19%	141	9%
Mpy.kandl1	1513	1391	913	59%	478	31%	151	10%
Mc.jannasc	1437	1338	883	57%	455	30%	204	13%
Fib.sucMM4	1354	1333	956	62%	377	24%	209	14%
Mccl.10mfx	1104	1100	787	51%	313	20%	442	29%

Table 3. Organisms for which articles have the least useful information. Each article is shown, along with the number of bases in *E. coli* for which it provides structural information. The next four columns show the average number of mismatches with the sequences (<IB>), the organism with the highest number of mismatches (Worst Org.), and the number (IB) and percentage (%IB) of bases in the article that have mismatches for this organism (%IB). The last four columns show the same statistics for gapped (II) bases-bases of *E. coli* for which there are no corresponding base in the tested sequence.

Article	# Bases	<IB>	Worst org.	IB	%IB	<II>	Worst org.	II	%II
jmb-moazed-187-399	1465	308.9	Msr.thmoph	490	33.4%	96.3	env.MC31	469	32.0%
gutell-ecoli-16s-sstruct	766	204.8	Pc.abyssi	318	41.5%	67.1	env.MC31	247	32.2%
rna-powers-1-194	592	105.0	Mcu.olenta	204	34.5%	19.7	env.MC31	183	30.9%
nar-stiege-1988-1	137	22.7	Msr.thmoph	47	34.3%	2.5	Spg.yanoi4	52	38.0%
jmb-powers-200-309	84	9.5	Msr.thmoph	22	26.2%	3.8	env.MC26	57	67.9%
jmb-stern-1988-201-683	79	8.2	Tpl.acidop	21	26.6%	2.8	Anbn.cylin	20	25.3%
jmb-svensson-200-301	66	14.2	Hb.sacchar	27	40.9%	0.3	Spg.yanoi4	11	16.7%
jmb-stern-200-291	53	5.6	Hb.sacchar	16	30.2%	1.6	env.OS_L	19	35.8%
jmb-stern-1986-1	39	7.2	Hb.sacchar	16	41.0%	0.8	Tms.thyasi	20	51.3%
nar-osswald-15-3241	39	7.0	Seca.cer_M	14	35.9%	10.9	env.MC31	23	59.0%
jmb-powers-201-697	37	3.5	Msr.thmoph	13	35.1%	0.8	env.MC31	25	67.6%
jmb-moazed-1990-1	18	0.5	Msc.aggreg	4	22.2%	3.0	Anbn.cylin	12	66.7%
nar-greuer-15-3221	18	1.7	Pyb.island	6	33.3%	0.3	env.MC26	14	77.8%
nar-camp-1989-1	16	2.0	Tpl.acidop	5	31.3%	2.5	Stv.luteor	9	56.3%
nar-brimacombe-6-1775	3	0.0	Tpl.acidop	1	33.3%	0.0	env.MC26	3	100.0%
sci-stern-244-783	32	5.2	Hb.sacchar	13	40.6%	0.0	Dgl.spRt46	3	9.4%

Figure 1. The web-based ONTOLINGUA browser shown with the hierarchy of data classes. ONTOLINGUA allows browsing and editing of this hierarchy, as well as the addition of instances. Part of the taxonomy of structural data is shown. Class names are indented to indicate the tree structure of the taxonomy—for example, `Cross-Linking-Data` is a type of `Biochemical-Data`.



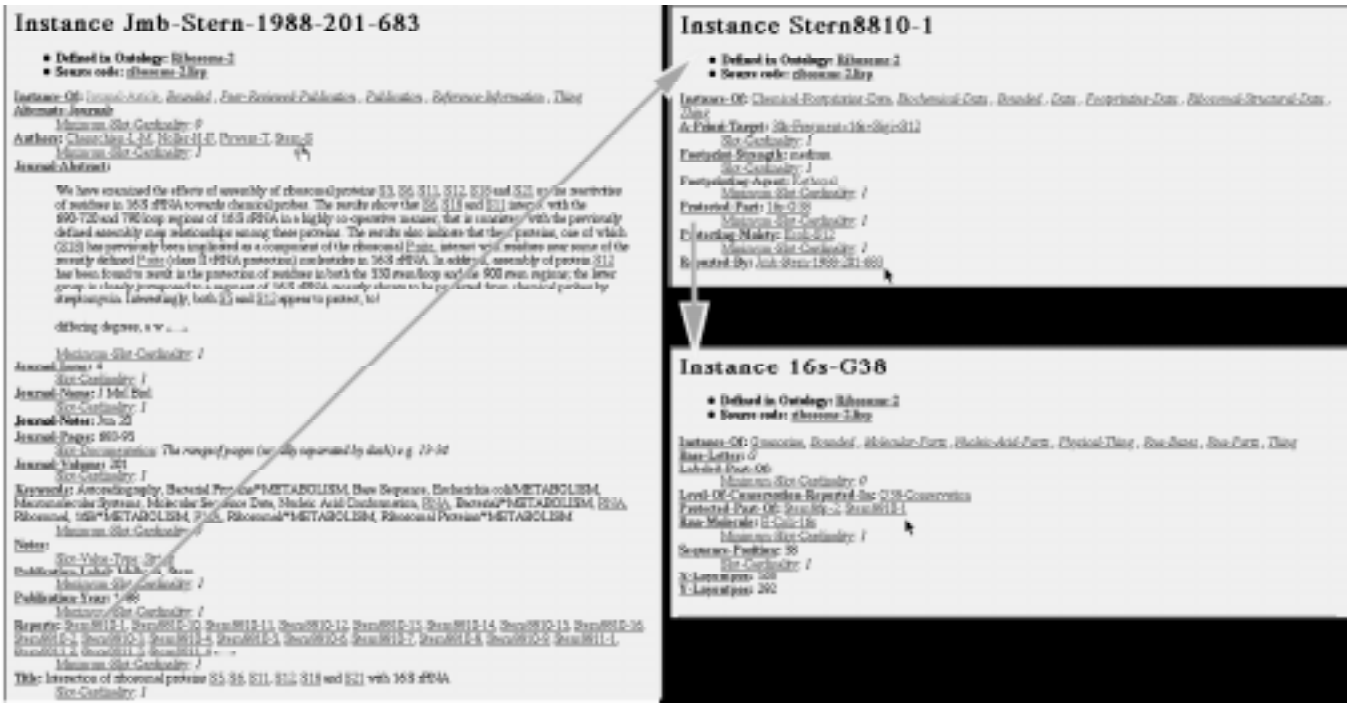


Figure 2. Three sample instances from the KB. (LEFT) The information for an article is shown, including bibliographic information and a list of reports links that point to standard representations of the data reported in the article. (TOP RIGHT) One of the data items reported by the article is shown, a chemical-footprinting-data that represents the information that protein S12 protects base G38 from attack by the small molecule kethoxal in *E. coli*. (BOTTOM RIGHT) The information for base G38 shows basic information about this bases, as well as links to other information that has been reported.

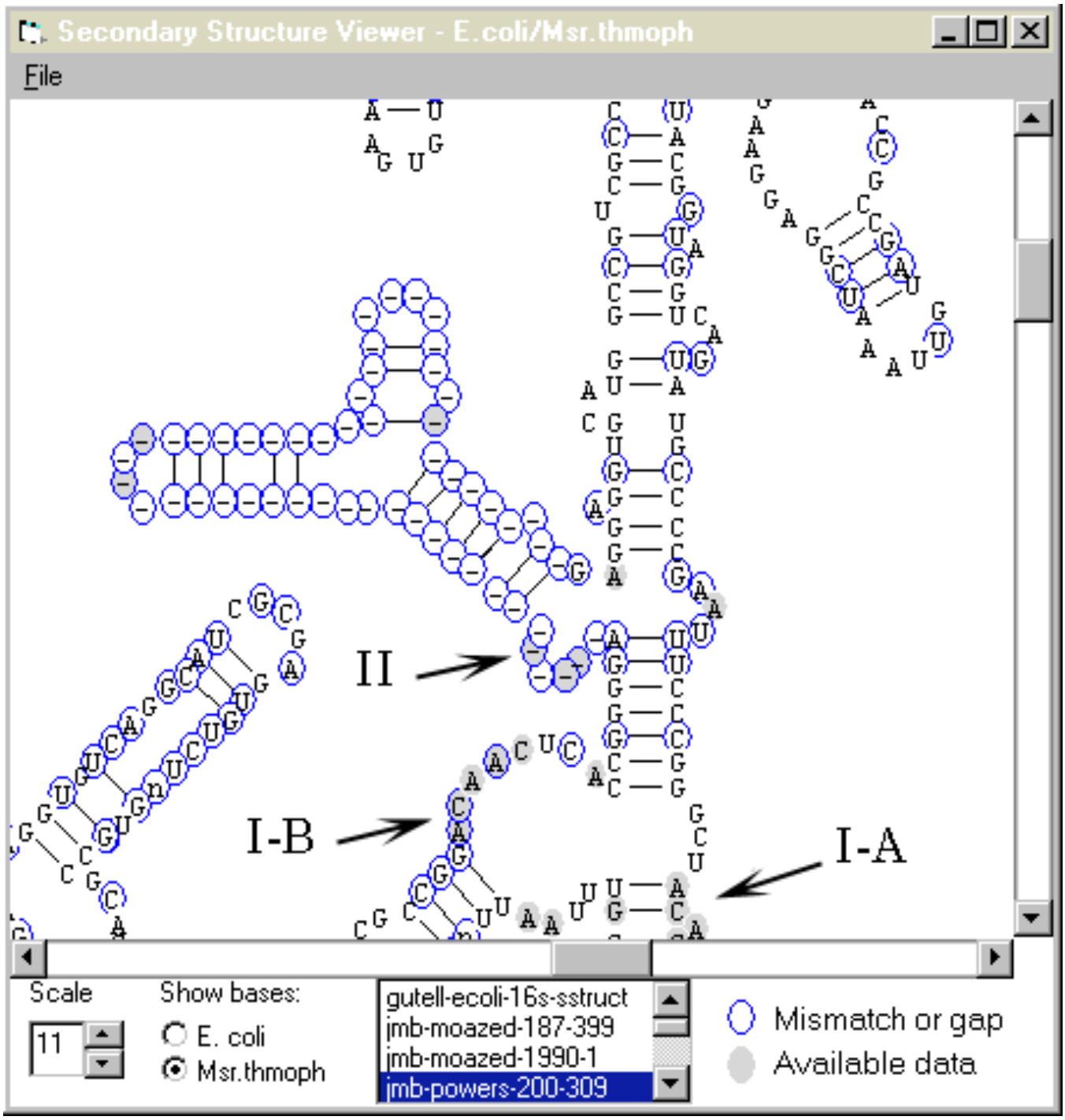


Figure 3. Comparison of *E. coli* sequence and secondary structure with Msr. thrmop, in the context of jmb-powers-200-309. This screen is created by our application. The *E. coli* secondary structure is shown with the corresponding bases from *Msr. thmoph*. A "-" is placed in positions where there is no base corresponding to a position in *E. coli*. For example, the large set of stem-loops in the upper left are not present in *Msr. thmoph* . The bases for which information is provided in jmb-powers-200-309 are shaded in gray. Sample relations of type IA (identity in alignment between *E. coli* and *Msr. thmoph* of a base mentioned in jmb-powers-200-309), IB (aligned mismatches between the two species), and II (a deletion in *Msr. thmoph* relative to *E. coli*) are shown with arrows.