8-2009

Evaluation of Annotation Performances between Automated and Curated Databases of *E.COLI* Using the Correlation Coefficient

ReddySalilaja Marpuri
Western Kentucky University, reddysailaja.marpuri418@wku.edu

EVALUATION OF ANNOTATION PERFORMANCES BETWEEN AUTOMATED AND CURATED DATABASES OF *E.COLI* USING THE CORRELATION COEFFICIENT

A Thesis
Presented to
The Faculty of the Department of Biology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirement of the Degree
Master of Science

By
ReddySailaja Marpuri

August 2009

EVALUATION OF ANNOTATION PERFORMANCES BETWEEN AUTOMATED AND CURATED DATABASES OF *E.COLI* USING THE CORRELATION COEFFICIENT

Date Recommended __July 31, 2009_____

__Claire A. Rinehart_____
Director of Thesis

__Sigrid Jacobshagen_____

__Cheryl Davis_____

_____
Dean, Graduate Studies and Research Date

Table of Contents

List of Figures

Page

List of Tables

EVALUATION OF ANNOTATION PERFORMANCES BETWEEN AUTOMATED AND CURATED DATABASES OF *E.COLI* USING THE CORRELATION COEFFICIENT

ReddySailaja Marpuri August 2009 51

Directed by: Claire A. Rinehart, Cheryl D. Davis and Sigrid Jacobshagen

Department of Biology Western Kentucky University

This project compared the performance of the correlation coefficient to show similarities in annotations between a predictive automated bacterial annotation database and the curated EcoCyc database. EcoCyc is a conservative multidimensional annotation system that is exclusively based on experimentally validated findings by over 15,000 publications. The automated annotation system, used in the comparison was BASys. It is often used as a first pass annotation tool that tries to add as many annotations as possible by drawing upon over 30 information sources. Gene ontology served as one basis of comparison between these databases because of the limited common terms in the ontology annotations. Translation libraries were used to extend the number of BASys terms that could be compared to the gene ontology terms in EcoCyc. Additional, non-ontology terms and metadata in BASys were compared to EcoCyc terms after parsing them into root words. The different term sources were quantitatively compared by using the correlation coefficient as the evaluation metric. The direct gene ontology comparison gave the lowest correlation coefficient. The addition of gene ontology terms to BASys by using translation tables of metadata greatly increased the correlation coefficient, which was comparable to the parsed word comparison. The combination of

enhanced gene ontology and parsed word methods gave the highest correlation coefficient of 0.16.

The controlled vocabulary system of gene ontology was not sufficient to compare two annotated databases. The addition of gene ontology terms from translation libraries greatly increased the performance of these comparisons. In general, as the number of comparison terms increased the correlation coefficient increased. Future comparisons should include the enhanced gene ontology dataset in order to monitor the organization pertaining to formal nomenclature and the datasets generated from Word parsing can be used to monitor the degree of additional terms might be incorporated with translation libraries.

Introduction

Genome sequencing and annotation are useful methods to understand the function and capacity of genomes. The rapid introduction of newly sequenced genomes provides a great challenge for annotators and annotation tools to reliably predict the functional properties of the genome. Genome annotation not only involves the identification of genes in terms of precise start and end sites but also describes the cellular components, molecular functions, and biological process in which a gene is involved. The increase in the volume of the genomic data has led to a greater demand for high throughput annotation methods. Most automated annotation techniques assign functions to newly sequenced genes based on similarity to previously annotated genes. This approach has a few problems. For example, if there was an annotation error in a previously annotated comparison genome, it will result in a whole family of mis-annotated genes being propagated into all subsequent genomes being derived from it. Automated annotation usually fails to meet the "gold standard" of the hand curated databases since the level of details in automated annotation systems is usually reduced, resulting in the classification of proteins into broader functional categories. To overcome this problem; ontology terms are used in automated databases as a means of understanding and recognizing the details of proteins to the level of curated databases.

As the process of annotation has matured, the scientific community continues to add new layers of analysis and interpretation to the genome annotations necessary to extract its biological significance. Annotation of a genome bridges the gap between the sequence and the biology of the organism. Annotation is helpful in identifying the key features of the genome, like genes and their products. Functional annotation is the

process of collecting information and describing a gene in terms of its molecular function, biological role, sub-cellular localization and expression pattern within an organism (1). Genome annotation is a multi-step process, which is engaged at three main levels: the nucleotide level, the protein level and the process level (2).

Nucleotide-level Annotation

Nucleotide-level annotation is a step where the raw genomic sequence is analyzed and forms the basis for subsequent levels of annotation. This step plays an important role in the identification of punctuation marks within the genome like genes, genetic markers and other landmarks like tRNAs, rRNAs, non-coding RNAs and repetitive elements. The principal activity of this step followed by 'gene finding' is to place all the landmarks into the genome such that they form a bridge between the genomic sequence and pre-existing genetic data. Nucleotide-level annotation also tries to identify any ancient duplications within the genome in an attempt to convert the raw DNA sequence into a set of easily recognizable landmarks and reference points. Using programs like BLASTN (3, U3) and SSAHA (4, U4), ePCR (5, U11) and BLAT (U13), one can perform the finding of these landmarks clearly and accurately. All of these factors play an important role in connecting pre-genomic research with that of the post-genomic research, such as attempting to identify genes involved in human diseases that have been previously studied through classical genetics. Gene finding is the first and foremost step in annotating a genome. In the case of prokaryotic genomes, gene finding is the process of identifying long open reading frames (6). However, in the case of eukaryotes the ambiguities associated with identifying these ORF's are known to increase with the length of the genome sequence. The reason behind these ambiguities is the signal-to-

noise ratio defined by the percentage of the genome that lies within the coding and non-coding regions. Additionally, the complexity in finding eukaryotic genes arises from the presence of splicing and alternative splicing regions.

Various gene prediction programs have been employed for eukaryotic genomes like GenScan (U5), Genie (U6), GeneMark.hmm (U7), Grail (U8), HEXON, MZEF (U21), Gene Finder (U31) and HMM Gene (U9). Each of these programs are known to have their own level of sensitivities but as the intergenic lengths of the genomes increase, the accuracy for all these gene identification programs was found to decrease rapidly (7). Instead of complete reliance on gene prediction programs, one of the most reliable ways of predicting a genome's organization depends upon aligning a region of the newly sequenced genome to a sequence in another genome that is already known to be transcribed. Matches between a new nucleotide sequence and a cDNA, EST or a gene match in another species are good examples of predicting genomes based on similarity. It is always important to note that the genomes are interspersed with various elements such as pseudogenes, non-coding RNAs, regulatory regions and repetitive elements.

The current models in gene prediction such as Grail/Exp (U8), Genie EST and Genome Scan (U5, 8) exploit the opportunity of combining *ab initio* gene prediction programs with sequence similarity data. The *ab initio* algorithms that were combined with similarity data outperformed those that did not take similarity results into account (9).

There is much more information encoded within a genome than just the protein coding regions. The cutting edge of nucleotide-level annotation is to search for

non-coding RNAs like tRNAs, rRNAs, small nuclear RNAs, small nucleolar RNAs and transcriptional regulatory regions. The development of algorithms to search for non-coding and regulatory regions in a genome has been one of the hot research topics in the field of bioinformatics. The prediction performances from these algorithms were found to be dependent upon the input sequences that were chosen as training sets.

Identification of repetitive elements is the next most important part in nucleotide annotation. Repetitive elements are one of the most dominant features of eukaryotic genomes and account for approximately 44% of nucleotides in the human genome (10). Repetitive elements are a nuisance for genome assembly and should be identified even before the annotation activities begin. Repeat masker (U14) is a program that identifies repetitive elements of a sequence. The identification of large segmental duplications was one of the biggest surprises that emerged during the sequencing of the mustard weed genome (11). Analysis of this sequence divergence has indicated four distinct segmental duplication events in the genome, spanning back to the period of angiosperm diversification (12).

The final activity of a nucleotide-level annotation is the identification and mapping of polymorphisms. The simplest way to identify single nucleotide polymorphisms, SNPs, is by aligning the genomic sequences of 2 or more genome sources and identifying where the sequence of one diverges from the other. Several algorithms have been developed to distinguish between true biological variations and differences due to errors in sequencing (4, 13, 14). Each of these methods use sequence quality data to make estimations. The distribution of SNPs into SNP hot spots was one of the intriguing findings from the annotation of human SNPs (2).

Protein level Annotation

Annotation at the protein level builds upon the identification of the coding ranges by translating the DNA sequence into amino acid sequence followed by the assignment of functions to the corresponding proteins. Annotators begin to classify and assign functions to new proteins based upon their sequence similarity to well-characterized proteins. A rich source for functional annotation comes by comparing proteins from different species. For example, if a well-characterized yeast protein is known to be involved in the process of DNA replication, then a human genomic sequence that is similar to the yeast protein is predicted to have the same functional outcome. To determine how valid the relationship is, the influence of evolution on the gene must be considered. The human protein might be a direct descendant of a common ancestor of yeast gene or it may be a descendant from a duplicated and diverged copy of the ancestor gene. In the latter situation, it is not correct to assume that the yeast and human proteins have the same functional role. It is common practice, however, to classify predicted proteins based on functional domains, folds and motifs and on similarity to well-characterized proteins. Various tools and databases like BLASTP (U16), SWISS-PROT TrEMBL, PFAM (U18), PRINTS (U20), ProDom (U23), Blocks (U26), SMART (U25) and InterPro (U27) have been used for this purpose.

Process-Level Annotation

The last and most challenging part of genome annotation is bridging the gap between the sequence and the biology of an organism. Functional annotation of a genome helps in evaluating known versus unknown genes in a genome. It is the relation of the genes and proteins to various processes of life like cell-cycle, cell death, embryogenesis,

metabolism, health and disease conditions that begin to define genes at the organismal level. The lack of a common classification scheme to describe a biological function among diverse species and to distinguish a particular protein from the other members of the family has hampered the ability to relate genes that were annotated by different research groups. A solution to this problem came through the consortium that developed the annotation schemes for three model organism databases, *Saccharomyces* Genome Database (15), Flyabase (U28,16) and Mouse Genome Database (U30,17) using a controlled vocabulary for Gene Ontology (18). A standard vocabulary for describing the function of eukaryotic genes is found in Gene Ontology (GO) (U29). "Controlled vocabularies are increasingly used by databases to describe genes and gene products because they facilitate identification of similar genes within an organism or among different organisms" (1). Gene Ontology (GO) has a tripartite structure consisting of: molecular function, biological process and sub-cellular localization.

Molecular function describes the task carried out by the product of a gene. Biological processes are defined as "a phenomenon marked by changes leading to a particular outcome, mediated by genes or gene products" (19). It has been observed that there is a close interrelation between molecular function and biological process. For example, the biological process named anti-apoptosis is involved in the molecular function known as apoptosis inhibition. This shows that the biological process is a collection of one or more molecular functions. Cellular component terms describe genes in terms of their sub-cellular localization. This framework of GO represents an anatomical counter-part, which allows biologists to know the physical structure within which the gene product is associated.

Each of the GO terms are organized in a hierarchical manner in such a way that the parent or broader concepts are on the top level of the tree structure and the more specific concepts, or child terms, are located on the branches. For example, a broad concept like enzyme leads to more specific terms like lyase, hydro-lyase and dehydrase. As new terms are added to the existing hierarchy the GO construction becomes more 'Bushy' (2). In conclusion, process level annotation extends well beyond computational work to a level that merges conventional genome research and genome annotation with biological context and function.

Model Organism for this Study: *E. coli* k12

E. coli k12 was originally isolated in 1922, and was known for its ability to carry out genetic recombination by conjugation (20) and by generalized transduction (21). It was developed into the primary model organism for prokaryotic biology, molecular genetics and physiology. The entire genome sequence of *E. coli* K12 MG1655 was determined and annotated by Fredrick Blattner *et al*. at the University of Wisconsin, Madison (22) in 1997. *E. coli* is an organism with a small genome size and much of the functional information is known. Annotation of the *E. coli* did not only serve the *E. coli* community but it was also helpful in extrapolating gene functions to both prokaryotic and eukaryotic genomes based upon similarity between protein sequences. Considering the degree of annotation completeness, the *E. coli* genome was used for this study to compare the automated bacterial annotation program of BASys with that of the manually curated and validated database of EcoCyc.

Manual curation

Manual curation involves extraction of functional information from the scientific literature to validate automated predictions and is found to be more accurate than automated annotations alone (23). The drawbacks associated with this technique are that it is time consuming and requires a skilled biologist to do it.

Automated Annotation

This process deals with comparing newly sequenced genomes to automated annotation databases and assigns functional information to the new genes based upon similarities in their sequence. Additional information is then required for validation from the literature. The success of automated annotation in predicting actual biological function comes in part from the strength of the protein sequence similarity. For example, two proteins having nearly identical sequences are more likely to have an identical function than two proteins with very divergent sequences.

Current Status

Genomes belonging to integrated databases can be viewed and interpreted as functional molecular data rather than as isolated sequence datasets. Initially, computational resources were developed to identify the features in the individual genes, such as their location, size, introns, promoters, and expressed sequence, but as the years have passed, a shift was observed towards the creation of annotation tools. These 'next-generation' comparative annotation tools have been used to create genomic databases and software packages that view and manipulate genome related information drawn from multiple sources. For example, annotations for a gene often include additional functional

information, such as enzymatic activity, substrates and their binding sites, cofactors, interactions with other proteins, regulation and timing of gene expression, alternate forms of the expressed gene, location of the protein within the cell and organism, the 3-dimensional structure of the protein, processing of the protein, polymorphisms within the population and relationships to disease.

Challenges associated with annotations:

The functions of newly sequenced genomes can be predicted initially by automated annotation methods. These methods usually start with a sequence comparison to databases that contain sets of genes for which annotations exist as, either validated or predicted. Based on the strength of sequence similarity, annotations are assigned from the associated gene matches. Therefore, the quality of the annotation information assigned to a new genome mainly depends upon the quality of the comparative annotations and the effectiveness of bioinformatics tools that assign them. Orphan genes, pseudogenes, and frame shifts pose a special challenge to automated annotation systems because they are either underrepresented in the comparative database or prove to be exceptions to the process.

The common challenge faced by most annotation systems is the need for rapid annotation of large amounts of sequence data when the data are riddled with inconsistencies between comparable genomes, like insertions and deletions due to recombination and transpositions. These inconsistencies often require human intervention to resolve conflicts and “finish” the annotations, thus slowing down the process. Several programs, such as MaGe, MiCheck, Phydbac and NMPDR, have used additional “context” or “synteny” information to try to resolve inconsistencies but even these

approaches are often limited by not enough comparable contextual data to facilitate a match. Therefore, it usually comes down to finding enough skilled eyes to update the databases and finish the sequence annotations.

The linking of data sources from various organizations is becoming a common theme in the annotation community, but this also poses a challenge to translate the disparate annotations for similar functions into a controlled vocabulary for gene ontology assignment. Several organizations such as JCVI, BioCyc, KEGG and NMPDR are striving to compile multiple genomes into large databases that use consistent and comparable nomenclatures. With the continuous increase in new genome sequences and functional annotation data, the great challenge is to integrate and distinguish validated vs. putative annotations while making all of the information available for rapid querying and retrieval. Accompanying this increased volume, there is a need for improved genome mining tools and controlled nomenclatures that will allow the correlatation of more data types into higher order relationships such as integrated metabolic and regulatory pathways.

Review of the Bacterial Annotation Systems

MICheck (U35): A web tool for fast checking of syntactic annotations of bacterial genomes.

A new web program MICHECK (Microbial Genome Checker), enables the rapid verification of annotated data sets and frame shifts in previously published bacterial genomes. The main emphasis of this program is to improve the status of annotations through re-assignment of hypothetical proteins to functional proteins based on

comparison to proteins with predicted functions. This is accomplished by identification of the coding sequences in the new target genome and all possible frame shifts that have shared synteny in other species. The annotations for the matching genes identified in this process are then assigned to the new target genome. Therefore, the MICHECK tool may improve the annotation quality of genome segments that have shared synteny with other genomes and can be used as a preliminary step to check for missing or wrongly annotated genes (24). However the information on pseudogenes is not usually represented in protein databanks resulting in the complete loss of these data.

MaGe (U34): Microbial Genome Annotation System supported by Synteny Results

MaGe, Magnifying Genomes, is a bacterial annotation system that identifies regions of conserved synteny between the target genome and other annotated genomes. Annotations are extrapolated to the target genome from these syntenic regions and the results are stored in a relational database (25). This database can be both queried and edited through a web interface that allows hand curation by several editors at once to refine assigned functions and classifications. It facilitates exploration of functionality by direct comparisons to reference data and incorporates the SEED (Subsystem) environment to identify uncalled genes.

AGMIAL (U32)

AGMIAL is a genome annotation system for prokaryotes that embodies a number of key principles used by expert manual curators. The philosophy underlying the Agmial platform is that the central role of the annotation process is in the hands of human experts

(26). This program is designed to assist annotators by providing databases to manage batches of assembled contigs and by providing automated visualization tools like those that identify sub-cellular localizations, multiple sequence alignments, or tools for finding motif profiles, regular expressions and protein family domains. In addition to these features a man-machine interface is provided for the annotators to interact efficiently with the data. Moreover, Agmial is found to be a modular system that facilitates integration of new tools. It is capable of handling draft sequences at various levels of completion and tracking the details of the changes occurring at the genomic or proteomic levels during the annotation process. Overall it provides a collaborative annotation process to annotate newly sequenced genomes and also provides good framework for re-annotation and data mining techniques. Because of its database, multiple genomes can be analyzed by the same set of tools and stored with the same type of organizations in Agmial. This facilitates a more accurate comparison between related organisms or strains.

DAVID (U36): Database for Annotation, Visualization and Integrated Discovery.

The functional annotation of differentially expressed genes is an important step in the analysis of microarray data. Annotation of a large number of differentially expressed genes can become burdensome without automated tools to extract relevant annotations from all of the major databases. The program named DAVID: Database for Annotation, Visualization, and Integrated Discovery, was designed to analyze large sets of genomic data and provide them with functional annotations (27). David is a web-accessible program that integrates functional genomic annotations with intuitive graphical

summaries. Currently, DAVID has over 40 annotation categories, including GO terms, protein-protein interactions, protein functional domains, disease associations, bio-pathways, sequence general features, homologies, gene functional summaries, gene tissue expressions, literature references, etc. It also has functional annotation clustering that helps measure the relationships between annotation terms on the basis of co-association thus helping annotators avoid the problem of associating different terms to the same biological process.

Phydbac: An interactive resource for annotation of bacterial genomes.

Phydbac uses phylogenomic profiling to annotate bacterial proteins. Phylogenomic profiling is based on the assumption that proteins involved in a common pathway might evolve in a correlated manner. The contextual evolutionary pattern of *E. coli* proteins can be identified by their similarity to query protein sequences based on the correlation with adjacent sequences. These patterns provide functional predictions for open reading frames with unknown or hypothetical functions. Additionally, the "Gene Function Predictor" of Phydbac provides translation between putatively identified *E. coli* K12 protein functions and GO term predictions (28).

SABIA (U33): A System for Automated Bacterial Integrated Annotation.

SABIA is a computational support system for the assembly and annotation of bacterial genomes. It streamlines the overall annotation process by connecting the assembly and annotation processes. SABIA uses a phred/phrap/consed package for the assembly of the contig data during the shotgun-sequencing phase (29). This connectivity between the sequencing and the assembly routines not only helps in streamlining the

overall process but also plays an important role in improving the quality of sequencing. SABIA identifies repetitive elements, ORF's, tRNA, protein motifs using bioinformatics tools like glimmer, BLAST, tRNA scan and InterPro. In addition to these features, rRNA sequences, frame shifts and missing steps in metabolic pathways are also identified.

NMPDR:

National Microbial Pathogen Database (30, U38) takes a subsystem approach for annotating microbial genomes. A subsystem is a set of functional roles that represent a specific biological process. In other words, it is like a pathway comprising a group of enzymes, each involved in a particular function (31). Initially the functional roles of a group of proteins are predicted. Then sequence similarity and gene co-localizations are used for consistent annotation. This is followed by the use of clustering analysis tools to identify genes involved in functions performed by divergent proteins, which cannot be recognized based on sequence similarity. Overall, these strategies make NMPDR a consistent functional annotation subsystem within a biological context.

KAAS (U39): An automatic genome annotation and pathway reconstruction server

KAAS is a KEGG Automated Annotation Server that performs automated functional annotations based on sequence similarity between target and annotated sequences by using BLAST for comparison of unknown targets to the manually curated KEGG genes database. KAAS works best with complete genomes because it can use bi-directional best-hit methods to define orthologs. Partial genomes and EST can also be annotated but uses the less stringent single-direction best-hit method. The KO (KEGG

Orthology) system is a metabolic pathway based definition for orthologous genes and was used by the KEGG database in annotating genomes. KEGG orthology is a controlled vocabulary like Gene Ontology that is directly linked to known metabolic pathways (32). KAAS was found to be a high-performance annotation tool for closely related organisms represented in KEGG genes database, however, it does not contain many closely related genomes to *E. coli.*

JCVI Automated Annotation Pipeline

The J. Craig Venter Institute has sequenced and annotated many genomes. This experience has led to the development of a pipeline of tools that assists in the manual curation of genes. The pipeline begins with a preliminary automated annotation that has a filter to remove uninformative annotations. Controlled vocabulary libraries are used to make functional assignments and to document the evidence supporting those assignments. Next a suite of web-based annotation tools allows the manual refinement and extension of the annotation assignments while using a consistent nomenclature (33).

EcoCyc (U1)

EcoCyc is an Encyclopedia for *Escherichia coli* K-12 genes and metabolism. It is a member of a large collection of pathway/genome databases called BioCyc (34). The EcoCyc database is a database for the model organism *E. coli* K-12, providing a structured symbolic representation of *E. coli* metabolic pathways, transport functions, and gene regulation. It is a database that has combined the ongoing literature curation of the *E. coli* genome with functional curation of its biochemical and genetic networks. With so much focus on its metabolic, biochemical, and genetic investigations, *E. coli* was found to be one of the best-studied organisms and a primary reference for extrapolating

functions to other organisms. The *E. coli* K-12 genome annotation in the EcoCyc database was found to be one of the most accurate and complete of all multi-dimensional genome annotation systems (35). Since EcoCyc has validated functions for a high fraction of the *E. coli* genes it has been found to be a resource for the analysis of other microbial genomes at the level of individual genes. EcoCyc has also been used to describe the sub-unit structures of many enzymes and therefore has been the source for training and validation datasets, and for algorithms that detect protein-protein interactions (36). The EcoCyc database provides not only a rich multi-dimensional annotation set for the *E. coli* genome, but also has a large complement of software tools for querying and visualizing the data. EcoCyc has become known as the living version of *E. coli* annotation with regular updates in gene sequences, functions and locations. EcoCyc annotations are validated by over 15,000 publications. “Bioinformatists use EcoCyc as a gold-standard data source for predicting functional relationships among proteins (37), operon predictors (38) and algorithms for predicting regulatory networks from gene expression data” (39). The underlying principle of the EcoCyc project was to store information for *E. coli* gene function and cellular networks using structured ontology terms that can be easily accessed by system biology computations.

BASys (U2)

BASys, also known as Bacterial Annotation Systems, is a web-served program that supports automated in-depth annotation of bacterial genomic sequences. It is built around about 30 programs describing approximately 60 annotation subfields for each gene. Some of the databases from which BASys derives its information are Pfam (40), Prosite (41), InterPro (42), HAMAP (43), TIGRfam (44) and PRINTS. The depth and the

details associated with the BASys exceed those found in a SwissProt entry. Various genome annotation tools like GeneQuiz (45), PEDANT (46), Genotator (47), MAGPIE (48) /BLUEJAY (49), GenDB (50) and the TIGR CMR are used by BASys (51) for the interpretation of genomic data. BASys is divided into three parts, each part being distributed between computing nodes to perform the annotations. The first component is the front-end web interface for submitting the raw genomic data, for scheduling and for monitoring the annotation progress. An annotation engine is the second component of BASys used for analyzing the chromosome data and for generating the annotations. The final component of BASys is the reporting system, which is used for representing the annotation output associated with each gene in the form of graphics, HTML or text.

The pipeline of the BASys annotation engine combines similarity searches to reference databases and model organisms with sequence and structural analysis methods to generate annotations. The reference databases associated with this annotation pipeline are UniProt (52) and CCDB (53). UniProt is a central database of protein sequences merging Swiss-Prot, TrEMBL and PIR-PSD to form the UniProt knowledgebase in order to provide a central source with annotations and functional information for comparison between databases. The Cyber Cell Database (CCDB) (U37) is a web-accessible relational database, which is supported by the International *E. coli* alliance. The main aim of this database is to stimulate the collection of *E. coli* validated annotations by providing web-based tools for correcting key information without duplicating existing resources.

Quantitative Comparison Metric

With the availability of an entirely validated database of annotations (EcoCyc) and an ambitious automated annotation system (BASys), the question of how similar the

annotations are between these two databases could be addressed. Currently, there appears to be no standard for quantitatively comparing the relatedness of the annotations between two databases. Therefore, the hypothesis of the present study was that a quantitative metric, such as the correlation coefficient, could be used to distinguish the performance of different annotation methods between the two databases. The correlation coefficient and the annotation term classification system from which it was calculated, (true positive, true negative, false positive and false negative), were also used to identify the performance of comparable annotation subsets between the two databases.

Materials and Methods

Databases

The main aim of this study was to quantitatively compare the results of manual and automated annotation programs. EcoCyc is a manually curated database for *E. coli* that only enters annotations that have been confirmed in the literature (35). BASys, known as Bacterial annotation systems, is an automated annotation system drawing information from different sources, including both curated and prediction databases (51). Therefore, EcoCyc is more conservative in its annotation than BASys, which seeks to add as many annotations as possible in order to direct bench top confirmation. Each of these two databases has set fields and the challenge in comparing the two databases is to be able to match relevant information between the database fields.

Database Term Extraction

Initial comparisons between databases can be done between correlated annotation fields in each database. Uncorrelated fields may require translation libraries to convert one database into terms that correlate with the other database. Uncorrelated terms may also be broken down into their rudimentary parts and their individual word elements can be compared. The annotation fields, described in each of the databases, are outlined in Table 1.

Table1. Database Correlated and Uncorrelated Fields. All term fields are shown for the EcoCyc database but only the fields used in the database comparisons are listed for the BASys database. The field names listed in A and B are correlated between the EcoCyc and BASys databases. The Gene Ontology Terms in the BASys database were converted to Gene Ontology Numbers by using a GOt2GOn translation library (54). The terms from the fields in D were also able to be translated into Gene Ontology Numbers by using specific Database2GO libraries (54). The terms in all of the listed fields, including C and E, were compared by parsing terms into words before comparisons were made.

<table>
<tr><th></th><th>EcoCyc</th><th>BASys</th></tr>
<tr><td>Correlated</td><td>A.
Gene Names
Gene Ontology Numbers</td><td>B.
Gene Names
Gene Ontology Terms, these were translated to Gene Ontology Numbers before comparison</td></tr>
<tr><td rowspan="2">Uncorrelated</td><td rowspan="2">C. (used word parsing)
EcoCyc
monomeric name
reference source
EC number
Swissport number
P, C and F
structural class
alternate gene name
InterPro
Pubmed Id
Go_Ref
Taxon number
product and date created</td><td>D. (used Database2GO translation tables)
PRINTS COG
EC number HAMAP
Pfam PROSITE
TIGRfam InterPro</td></tr>
<tr><td>E. (used word parsing)
Protein_name EMBL
Alternate protein name PIR
specific-function ProDom
Swiss-2D PAGE substrates
operon component databases
Structural class similarity
general reaction cell location
specific reaction
inhibitor and references</td></tr>
</table>

Conversion of terms and database elements into comparable GO numbers.

The direct comparison of the correlated Gene Ontology (GO) fields between the EcoCyc and BASys databases first required the translation of the BASys Gene Ontology (GO) terms (Table 1B.) into Gene Ontology (GO) numbers similar to Table 1A. The February 6, 2009 version of the GO number to GO term translation table,

"GO.terms_and_ids", from the Gene Ontology site (54) contained 80038 number/term combinations that allowed translation of the BASys GO terms into GO numbers. Additional translation tables were available from the Gene Ontology site (54) to convert terms, from the eight database derived fields listed in Table 1C. into GO numbers. A flow diagram of all of these translation processes along with the specific tables used is shown in Figure 1. These translation tables are updated on varying schedules and the tables that were used in the BASys field translations were downloaded in February 2009.

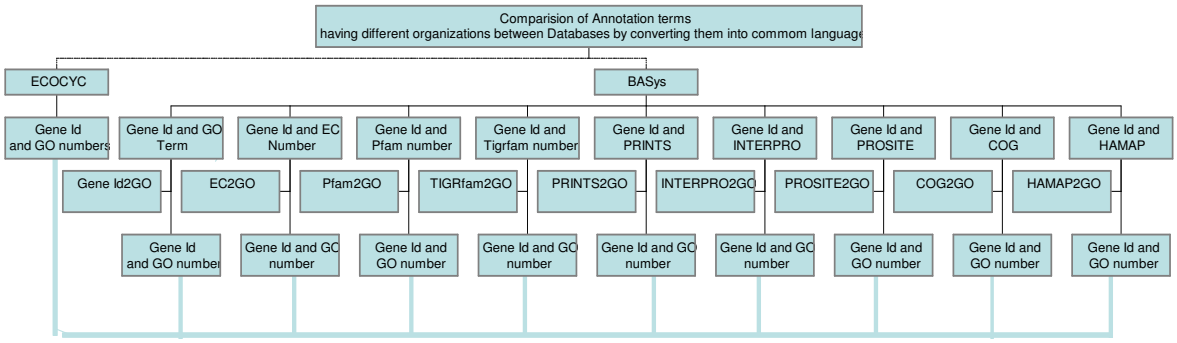


Figure 1. Flow chart for translation of annotation terms and database elements into Gene Ontology numbers. The Gene Ontology numbers for each gene ID in EcoCyc was compared to Gene Ontology numbers for each gene ID in the BASys database (bottom row of boxes under BASys). The Gene Ontology numbers for the BASys dataset were translated from annotation terms (top row of boxes under BASys) by using database-specific translation tables (middle row of boxes under BASys).

Conversion of terms and database elements into comparable text words and dictionaries.

Another approach for comparison of terms between the databases was through text mining (55, 56, 57). The first step in text mining was to break the text into comparable chunks or words. The annotation data associated with common gene IDs in both the databases was extracted and parsed out into word terms by splitting the text at all

of the delimiters: : ; , . / { } ( ) [ ] $ • ! @ % * _ + = \ > ” ? | & space. This created comparable sets of word terms for each gene ID in EcoCyc and BASys.

Database Term Comparisons

The translation and parsing methods described above created comparable sets of terms for the gene IDs in EcoCyc and BASys. To determine how many of these terms were common between the two databases, a non-redundant dictionary, with terms common to both EcoCyc and BASys, was created from the intersection of the database's set of comparable gene IDs. These comparable terms were in the form of either GO numbers or parsed words. Once the data were converted, the information was compared between the databases.

Description of the comparison process:

1. Initially BASys and EcoCyc were compared against each other to find the common gene IDs between them and a common ID dictionary was created (cid).
2. The records associated with common IDs for each database were extracted, and translated or parsed into a format of comparable terms (term books). The number of terms in each term books (E@cid or B@cid) was totaled.
3. The dictionary of terms common to both the term books was created from the intersection of the terms found in the two books (Dict).
4. To determine how many of the terms in each database's term books had the possibility of being matched to the other database, the terms of each cid were intersected with the dictionary of common terms and these terms were kept

(E2Dict@cid or B2Dict@cid). The numbers of these dictionary-intersected terms were totaled.

5. Finally, the terms at each cid were compared between E2Dict@cid and B2Dict@cid and those that were found to be common were saved as true positive matches (E2B4Dict@cid). The numbers of true positive matches were totaled.

Table 2. Names for totals resulting from database term extraction process. Columns 2 and 3 correspond to the EcoCyc and BASys extracted datasets respectively and the names contain E and B as designators of the database. The @cid part of the name indicates that the extracted sets were derived from each record that had a gene name in common between the databases. The 2 in the name indicates the intersection between 2 datasets. Dict refers to the dictionary of common terms between the two databases. The 4Dict following the E2B indicates that the intersection was between the dictionary qualified sets E2Dict and B2Dict at each cid.

Database	EcoCyc	BASys
Extract all terms from records with common IDs	E@ cid	B@ cid
Additionally, extract terms common to Dictionary at each id	E2Dict@ cid	B2Dict@ cid
Extract terms common between databases at each id	E2B4Dict@cid	

Quantifying Database Relationships.

The EcoCyc database was taken as the most accurate since the annotations have been verified from the literature. The correlation coefficient (CC), defined by formula 1, was used as the metric to compare the performance of the BASys annotation predictions to the curated EcoCyc annotations. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), were defined by formulas 2 through 5 respectively.

$$CC = [(TP * TN) - (FP*FN)] / \sqrt{[(TP+FN)(TN+FP)(TP+FP)(TN+FN)]} \quad (1)$$

$$TP = E2B4Dict@\ cid \quad (2)$$

$$TN = (E@\ cid - E2Dict@\ cid) + (B@\ cid - B2Dict@\ cid) \quad (3)$$

$$FP = B2Dict@\ cid - E2B4Dict@\ cid \quad (4)$$

$$FN = E2Dict@\ cid - E2B4Dict@\ cid \quad (5)$$

The True Positives are those terms that match between records with corresponding gene names. The False Positives are those annotation terms that were predicted by BASys but not validated in EcoCyc. The False Negatives are those annotation terms that have been validated in EcoCyc but were not identified by the BASys annotation program. The True Negatives are the large class of annotations that are not found in the term dictionary (Dict) and therefore are not comparable terms between the two databases.

Results

The goal of this study was to quantitatively compare how well annotations for the *E. coli* genome matched between the manually curated and automated databases. The model organism used in this study was the *E. coli* bacterium, because it is one of the best-studied primary reference organisms. BASys is an ambitious, automated annotation system that draws its annotations from many other databases. EcoCyc is a manually curated database that requires its annotations to be validated from the literature before they are entered into the database. Comparison of the BASys database to the EcoCyc standard reveals how well the BASys automated annotation system performs in its annotation predictions.

In order for the genomes to be compared there must be a common factor between the databases that has all the annotation information (fields) associated with it. The gene IDs were initially taken as the common factor for comparison of the databases. EcoCyc has about 15 annotation fields associated with each of its gene IDs, whereas BASys has about 56 of them associated with each gene ID. EcoCyc has 1594 unique gene IDs in 4712 annotation records and BASys has 4254 record/gene IDs out of which 1335 gene IDs were common between both the databases. Only records that had common gene IDs (cid) were taken into consideration for comparison. There exists some redundancy in the EcoCyc ID data because multiple records with the same gene ID were used to extend the number of annotations. The advantage of considering only the records with common gene IDs helps not only in eliminating records that cannot be compared but also in preventing the generation of huge volumes of true negative data. This was important, since the goal of this study was to quantitatively determine how well BASys modeled the true set of

annotations found in EcoCyc. The records associated with common gene IDs were placed into the E@cid and the B@cid datasets respectively from the EcoCyc and BASys databases (See Material and Methods).

A dictionary of common terms, either GO numbers or parsed words, was created by taking the intersection of E@cid and B@cid for the respective terms following extraction, translation and parsing of the original records (See Materials and Methods, Conversions…). The dictionary of common terms was compared to each of the term datasets, E@cid and B@cid on a record-by-record basis to define the set of words in each record that had the possibility of matching between the databases. These terms were saved in the datasets E2Dict@cid and B2Dict@cid for the EcoCyc and BASys records respectively. Finally, the terms in each corresponding record of E2Dict@cid and B2Dict@cid were compared, and terms common to both records were saved into the E2B4Dict@cid dataset as the true positive matching terms. From these datasets the number of true positive matches, true negative matches, and the number of false positives and false negatives were totaled and used to calculate the correlation coefficient between the two databases for each set of terms being considered (See Materials and Methods, Database Term Comparisons).

Term Comparisons

Four sets of term comparisons were made: 1) A direct comparison of gene ontology numbers, 2) A comparison of gene ontology numbers after translation of other field terms into gene ontology numbers, 3) A comparison of parsed words from all the fields in each database, and 4) A comparison of terms combined from 2 and 3 (See

Materials and Methods, Table 1 for an explanation of database fields used in each comparison). The results from these comparisons are presented below.

1) Direct Comparison of Gene Ontology Numbers.

Out of the total 80038 GO numbers in the gene ontology site (54) 7705 of them were used to convert BASys GO terms to GO numbers. It was found that about 44% percent of the BASys GO terms were translated into GO numbers using the conversion tables Got2GOn in gene ontology site (54). The results for each of the extracted term classes are shown in Table 3. For both the databases, approximately 50% of the gene ontology number terms were found in the common term dictionary (Table 3., line 2 / line1). For the EcoCyc terms common to the term dictionary, only 22% matched terms in their respective genes in BASys (E2B4Dict@cid / E2Dict@cid). Only 9% of the BASys term common to the dictionary successfully matched terms in their respective genes in EcoCyc (E2B4Dict@cid / B2Dict@cid).

Table 3. Comparison of Gene Ontology terms between databases. The numbers in parenthesis represent the total number of gene ontology numbers represented in each dataset. E@cid and B@cid are the datasets of all terms with common gene IDs for the EcoCyc (E) and BASys (B) databases respectively. E2Dict@cid and B2Dict@cid are sets of terms at each common ID (cid) that match a dictionary of common terms (Dict). E2B4Dict@cid is the set of terms common between E2Dict@cid and B2Dict@cid at each cid.

E@cid (3060)	B@cid (7705)
E2Dict@cid (1556)	B2Dict@cid (3838)
E2B4Dict@cid (349)	

2) Direct Comparison of Gene Ontology Numbers after translation of additional terms.

In the next step enhanced Gene Ontology terms were compared between the databases. Enhanced GO terms were obtained by converting some of the uncorrelated records associated with the common gene IDs from BASys into GO numbers. These conversions for the BASys records were made using the translation information in the db2GO files in the Gene Ontology site (Table 2., materials and methods). The resulting GO numbers from BASys were then compared to EcoCyc GO numbers and the results are shown in Table 4. For the EcoCyc database, 2473/3060 = 81% of the gene ontology number terms were found in the common term dictionary (Table 4.) while the BASys database showed a 6830/31641 = 21% match to the term dictionary. For the EcoCyc terms common to the term dictionary, only 838/2473 = 34% matched terms in their respective genes in BASys (Table 4.). Only 838/6830 = 12% of the BASys terms common to the dictionary successfully matched terms in their respective genes in EcoCyc (Table 4.).

Table 4. Comparison of enhanced Gene Ontology terms between databases. The numbers in parenthesis represent the total number of extended gene ontology numbers represented in each dataset. E@cid and B@cid are the datasets of all terms with common gene IDs for the EcoCyc (E) and BASys (B) databases respectively. E2Dict@cid and B2Dict@cid are sets of terms at each common ID (cid) that match a dictionary of common terms (Dict). E2B4Dict@cid is the set of terms common between E2Dict@cid and B2Dict@cid at each cid.

E@cid (3060)	B@cid (31641)
E2Dict@cid (2473)	B2Dict@cid (6830)
E2B4Dict@cid (838)	

3) Direct Comparison of Parsed Word Terms.

The results from the above two comparisons indicated that the percentage of relatedness between the databases was not high and that there were a lot of terms that were not being compared. This was because there were no more translation dictionaries for the conversion of the remaining terms. Therefore a data mining technique was applied that took all of the record information for common gene IDs, extracted and parsed them out into individual word terms to make term books for each database (see Database term extraction, materials and methods). These word (term) books from the individual databases were then compared and the results are shown in Table 5.

Table: 5 Comparison of parsed Word terms between databases. The numbers in parenthesis represent the total number of parsed word terms represented in each dataset. E@cid and B@cid are the datasets of all terms with common gene IDs for the EcoCyc (E) and BASys (B) databases respectively. E2Dict@cid and B2Dict@cid are sets of terms at each common ID (cid) that match a dictionary of common terms (Dict). E2B4Dict@cid is the set of terms common between E2Dict@cid and B2Dict@cid at each cid.

E@cid (29995)	B@cid (114653)
E2Dict@cid (14482)	B2Dict@cid (51863)
E2B4Dict@cid (7718)	

The results for each of the parsed word term classes are shown in Table 5. For the EcoCyc database, 14482/29995 = 48% of the parsed word terms were found in the common term dictionary (Table 5.) while the BASys database showed a 51863/114653 = 45% matched the term dictionary. For the EcoCyc terms common to the term dictionary, only 7718/14482= 53% matched terms in their respective BASys genes (Table 5.). Only 7718/51863= 15% of the BASys terms common to the dictionary successfully matched terms in their respective EcoCyc genes (Table 5.).

4) Direct Comparison of Parsed Word Terms combined with Extended Gene Ontology Terms.

Once the above comparisons were done, the extracted terms from the word parsing method and the enhanced gene ontology terms were combined. The combined term books were compared between both the databases and the results are shown in Table 6.

Table: 6 Comparison of combined Word terms and Enhanced Gene Ontology numbers between databases. The numbers in parenthesis represent the total number of parsed word and gene ontology number terms represented in each dataset. E@cid and B@cid are the datasets of all terms with common gene IDs for the EcoCyc (E) and BASys (B) databases respectively. E2Dict@cid and B2Dict@cid are sets of terms at each common ID (cid) that match a dictionary of common terms (Dict). E2B4Dict@cid is the set of terms common between E2Dict@cid and B2Dict@cid at each cid.

E@cid (29995)	B@cid (132035)
E2Dict@cid (16648)	B2Dict@cid (56494)
E2B4Dict@cid (9890)	

For the EcoCyc database, 16648/29995= 56% of the parsed word terms were found in the common term dictionary (Table 6.) while the BASys database showed a 56494/132035= 43% match to the term dictionary. For the EcoCyc terms common to the term dictionary, only 9890/16648= 59% matched terms in their respective BASys genes (Table 6.). Only 9890/56494= 18% of the BASys terms common to the dictionary successfully matched terms in their respective EcoCyc genes (Table 6.).

Quantitative analysis of the four comparison methods.

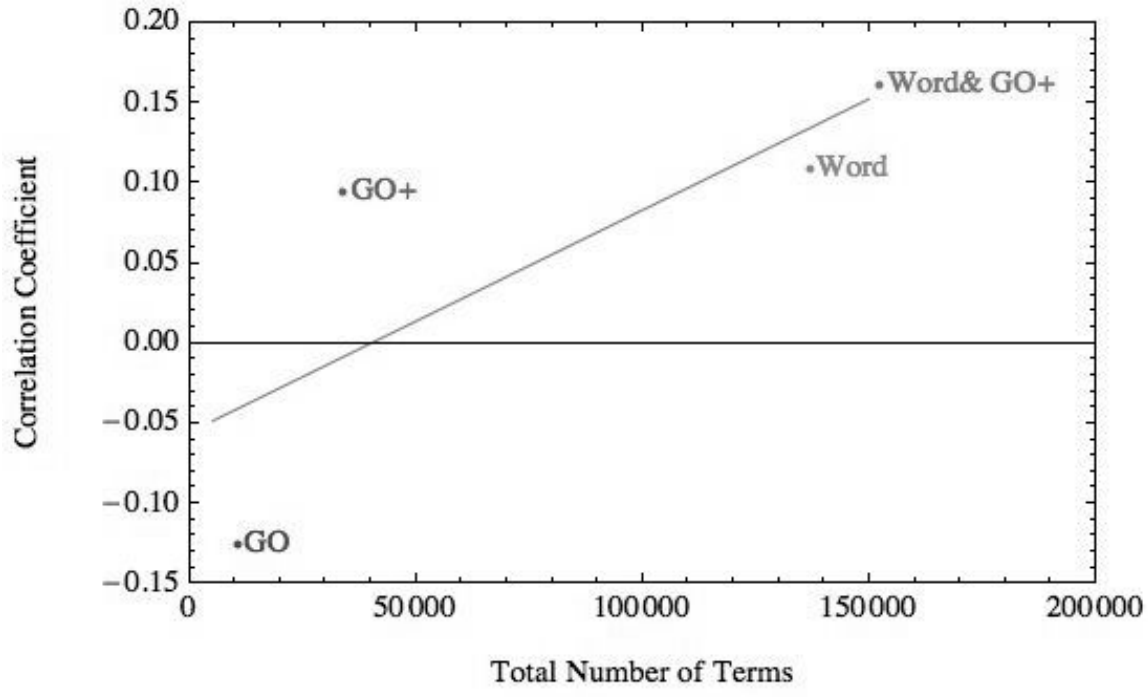
The correlation coefficient, defined by formula 1 (Materials and Methods), was used as the metric to compare the performance of the BASys annotation predictions based on the curated EcoCyc annotations. The true positives, true negatives, false positives, and false negatives defined by formulas 2 through 5 respectively (Materials and Methods) formed the basis for calculating the correlation coefficient. The numeric totals for each of these results for each of these performance classes are summarized in Table 7.

Table: 7 Quantitative Comparison of term extraction methods. The four term extraction methods are Gene Ontology (GO), Enhanced Gene Ontology (GO+), parsed word (Word) and the combination of the parsed word and enhanced gene ontology (Word&GO+). The true positives (TP) represent the total number of terms matching between the databases for common IDs. The false positives (FP) represent the total number of common dictionary terms in BASys that do not match their respective gene ID in EcoCyc. The true negatives (TN) represent the total number of terms in EcoCyc and BASys that are not found in the common term dictionary. The false negatives (FN) represent the total number of common dictionary terms in EcoCyc that do not match their respective gene ID in BASys. The correlation coefficient (CC) is the quantitative metric used to compare the extraction methods and is calculated from the formula:
CC = [(TP * TN) – (FP*FN)]/ Sqrt [(TP+FN) (TN+FP) (TP+FP) (TN+FN)].

Term Extraction Methods	True Positives	False Positives	True Negatives	False Negatives	Total	Correlation Coefficient
GO	349	3489	5371	1207	10416	- (0.125)
GO+	838	25398	5992	1635	33863	0.095
Word	7718	44145	78303	6764	136930	0.109
Word & GO+	9890	88888	46604	6758	152140	0.161

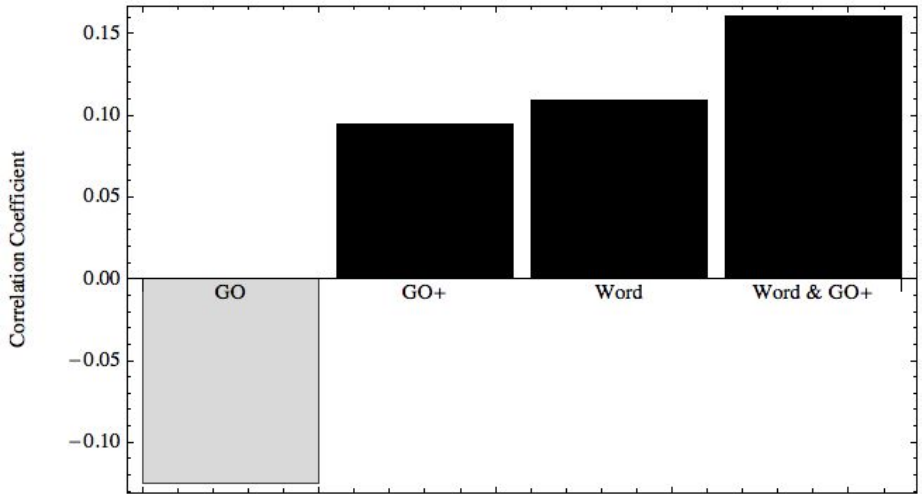
The correlation coefficient generally increased as more total terms were added by the different extraction methods (Figure 2.). The greatest performance gains in the correlation coefficient came by using translation libraries to convert the associated database terms into the GO+ numbers (Figure 2). On a total term basis, this was much more efficient than using the Word comparison methods which more than tripled the number of terms in order to gain a slight increase in correlation coefficient performance over the GO+ method (Figure 2).

Figure 2. Linear relationship between the number of terms considered and the correlation coefficient. The four term extraction methods are Gene Ontology (GO), Enhanced Gene Ontology (GO+), parsed word (Word) and the combination of the parsed word and enhanced gene ontology (Word & GO+). The line represents the best fit to the correlation coefficient and total number of terms data. The R^2 for the fit model is 0.62.



When terms from the enhanced gene ontology (GO+) were added to the parsed word terms (Word) the performance gain was not additive (compare the sum of GO+ and Word correlation coefficient to Word & GO+ coefficient value in Figure 3). This indicates that there was a fair amount of overlap between the two methods as indicated by the small increase in total considered terms between the Word and Word & GO+ values (Figure 2).

Figure 3. Correlation coefficient for each extraction method. The four term extraction methods were Gene Ontology (GO), Enhanced Gene Ontology (GO+), parsed word (Word) and the combination of the parsed word and enhanced gene ontology (Word & GO+).



DISCUSSION

The goal of our study was to compare automated annotation methods to a curated reference database in order to evaluate their performance by using the correlation coefficient metric.

The main purpose for annotation comparisons between genomes was to evaluate the accuracy of predicted functions associated with newly sequenced genomes. In doing such comparisons it is always important that the reference database should have annotations that are validated through scientific literature in the publications. Therefore it is always important to use some caution in using databases, as references, that are derived by automated annotation from unvalidated sources. The databases chosen for this study were BASys (automated) and EcoCyc (manually curated and literature validated).

The correlation coefficient is a performance metric that reflects the relative number of terms in four different comparison classes: true positives, false positives, true negatives, and false negatives (equation 1). The terms that belong to the true positive class are those terms that not only are found in the two databases (dictionary terms) but also match to a specific gene in both databases. This means that for a particular gene these matching terms are both validated in EcoCyc and predicted by BASys. Terms in the false negative class are also found in the dictionary and validated in EcoCyc but are not present in the corresponding gene in BASys. This means that the BASys annotation methods failed to identify this corresponding term. There may be two reasons for this failure, first, BASys may have predicted the correct function and called it by another name or second, BASys may not have identified the function at all. If a common function was called by another name, term translation tables that convert BASys terms to

the corresponding EcoCyc terms will greatly reduce the number of false negatives. Another class that benefits from the use of translation tables is the true negative class. This class represents the terms that are found in only one of the two databases and are not included in the common dictionary of terms. Translation of BASys terms into their corresponding EcoCyc terms increases the size of the dictionary and the number of terms considered for matching at each common gene id. This class is often large, especially when comparing databases with non-standard or proprietary annotation terms. The very reason for developing gene ontology methods was to encourage the use of a controlled vocabulary that could be easily compared between databases and reduce the size of this class. Finally, the false positive class of terms are found in BASys and the common term dictionary but do not match the corresponding gene id in EcoCyc. There are two reasons that the BASys terms may not match the corresponding gene in EcoCyc, first, the function identified in BASys may be correct but has not been validated and therefore is not found in EcoCyc, or second, BASys has incorrectly identified the associated function. The size of the EcoCyc database is continuously growing as new gene function validations become available for E. coli, therefore, increasing the possibility that "unvalidated" gene functions will be moved to the true positive class.

The question addressed by this study was, how well automated annotation methods compare to hand curated and validated annotation methods where with the correlation coefficient used to evaluate performance. In order to answer this question, comparable sets of data needed to be extracted and compared from each database. Four extraction methods were compared to each other and a summary of the results is found in Table 8. The simplest extraction method was to directly compare the Gene Ontology,

GO, data sets. This resulted in a negative correlation coefficient and 3% of the total terms showing a positive match between the databases (Table 8). One of the biggest factors in this poor performance was the large false negative pool of 12% (Table 8). The false negatives are terms validated in EcoCyc that did not have matching values in BASys. The numbers of false negatives were greatly reduced, to 5% (Table 8), by translating other database terms found in BASys to GO+ numbers. This resulted in a 0.22 point increase in the correlation coefficient even though the true positives dropped from 3% to 2% of the total terms. Note the large decrease in true negatives as the translation of BASys terms increased the common dictionary size and provided more comparable terms. Most of these additional terms ended up in the false positive class and may be validated at some future time with further hand curation of EcoCyc.

When comparing EcoCyc and BASys only a small number of information fields were used for gene ontology (Table 1, A & B for GO and A, B, and D for GO+), many additional terms were available for comparison (Table 1, C & E). Therefore a data mining approach was taken which reduced all of the terms listed in Table 1, A through E, into simple word terms. This approach gave a correlation coefficient performance that was slightly higher than the GO+ with a significant increase in the percent of true positives and a comparable percentage of false negatives (Table 8). As in the transition from GO to GO+, the addition of the GO translation tables (Table 8, Word & GO+) to the Word dataset (Table 8) greatly reduced the percent of true negatives, from 57% to 31%. The application of GO+ translations reduced the percent of false negatives as the additional BASys terms were made available to match the EcoCyc terms. The use of these translation tables resulted in over a 50% gain in the magnitude of the correlation

coefficient (Table 8). The application of additional translation tables, such as a synonym library, should result in additional performance gains.

Table: 8 Percent Comparison of term extraction methods. The four term extraction methods are Gene Ontology (GO), Enhanced Gene Ontology (GO+), parsed word (Word) and the combination of the parsed word and enhanced gene ontology (Word & GO+). The numbers for the true positives (TP), the false positives (FP), the true negatives (TN), and the false negatives (FN) represent the percent of the total terms represented in each class. The correlation coefficient (CC) is the quantitative metric used to compare the extraction methods and is calculated from the formula:

CC = [(TP * TN) – (FP*FN)]/ Sqrt [(TP+FN) (TN+FP) (TP+FP) (TN+FN)].

Term Extraction Methods	True Positives	False Positives	True Negatives	False Negatives	Correlation Coefficient
GO	3%	33%	52%	12%	- (0.125)
GO+	2%	75%	18%	5%	0.095
Word	6%	32%	57%	5%	0.109
Word & GO+	7%	58%	31%	4%	0.161

The conclusions that can be drawn from this study are that the controlled vocabulary system of gene ontology is not sufficient to compare two annotated databases. The addition of gene ontology terms from translation libraries greatly increased the performance of these comparisons. Term selection methods that increased the total number of terms being considered by the analysis generally increased the correlation coefficient. The combination of the enhanced gene ontology and parsed word methods not only resulted in an increase in the number of terms but also gave the highest correlation coefficient. The text-mining approach of parsing the annotations into words increased the number of comparable terms but resulted in the loss of context, which is an

inherent property of gene ontology. Therefore, future comparisons should include the GO+ in order to monitor the organization pertaining to formal nomenclature and the Word parsing to monitor the degree of additional terms that could be incorporated with translation libraries.

The correlation coefficient is a good single metric to define performance of not only the term extraction methods but also different automated annotation systems. The four classes for term characterization (true positive, false positive, true negative and false negative) also served as indicators of where performance gains can be made. For example, high false negative and true negative classes probably need additional term translation tables. By using the correlation coefficient, the Word and GO+ term extraction methods were sufficient approaches to determine the performance of automated annotation systems.

Currently only 1594 of the almost 4200 genes from *E. coli* have validated annotations in EcoCyc and much growth is anticipated in future updates. Therefore, in order to compare new annotation systems to those currently in use, the analysis of all comparable systems must be analyzed at the same time with the latest EcoCyc database values.

References

1. Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D. and Rhee. S.Y. (2004). Functional Annotation of the Arabidopsis Genome Using Controlled Vocabularies. Plant Physiology 135: 745-755.
2. Stein, L. (2001). Genome Annotation: From Sequence To Biology Nature Reviews 2, 493-503.
3. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215: 403–410.
4. Ning, Z., Cox, A.J. and Mullikin, J.C. (2001). SSAHA: A fast search method for large DNA databases. Genome Res. 11: 1725-1729.
5. Schuler. G.D. (1997). Sequence mapping by electronic PCR. Genome Res. 7: 541–550.
6. Field, D., Fiel E.J. and Wilson. G.A. (2005). Databases and Software for the comparison of prokaryotic genomes. Microbiology 151: 2125-2132.
7. Guigo, R., Agarwal, P., Abril, J.F., Burset, M. and Fickett, J.W. (2001). An assessment of gene prediction accuracy in large DNA sequences. Genome Res. 10: 1631–1642.
8. Yeh, R-F., Lim, L.P. and Burge, C.B. (2001). Computational inference of homologous gene structures in the human genome. Genome Res. 11: 803-816

9. Reese, M.G., Hartzell, G., Harris, N. L., Ohler, U., Abril, J.F. and Lewis, S.E. (2000). Genome annotation assessment in Drosophila melanogaster. Genome Res. 10, 483– 501.

10. International Human Genome Sequencing Consortium (IHGSC). (2001). Initial sequencing and analysis of the human genome. Nature 409: 860–921.

11. Arabidopsis Genomics Initiative (AGI). (2000). Analysis of the Arabidopsis Genomics Initiative (AGI). Analysis of the thaliana. Nature 408, 796–815.

12. Ku, H.M., Vision, T., Liu, J. and Tanksley, S.D. (2000). Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc. Natl Acad. Sci. 97: 9121–9126.

13. The SNP Consortium. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933.

14. Marth, G.T., Korf, I, Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri. H., Stitziel, N.O., Hillier, L., Kwok, P.Y., and Gish, W.R. (1999). A general approach to single-nucleotide polymorphism discovery. Nature Genet. 23: 452–456.

15. Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. and Botstein, D. (1998). SGD: Saccharomyces Genome Database. Nucleic Acids Res. 26: 73–79.

16. The FlyBase Consortium. (1999). The FlyBase database of the Drosophila Genome Projects and community literature Nucleic Acids Res. 27: 85–88.

17. Blake, J.A., Eppig, J.T., Richardson, J.E., Bult, C.J. and Kadin, J. A. (2001). The Mouse Genome Database (MGD): integration nexus for the laboratory mouse. Nucleic Acids Res. 29: 91–94.

18. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. Nature Genet.25: 25–29.

19. Smith, B., Williams, J. and Schulze-Kremer, S., (2003). The Ontology of the Gene Ontology. Proceedings of AMIA Symposium.

20. Lederberg, J. and Tautum, E.L. (1946). Gene recombination in Escherichia coli. Nature: 158: 558.

21. Lennox, E.S. (1955). Transduction of linked genetic characters of the host by bacteriophage P1. Virology 1: 190-206.

22. Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997). The Complete genome sequence of Escherichia coli K-12. Science 277:1453-74.

23. Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A. and Apweiler, R. (2003). The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL and InterPro. Genome Res. 13: 662-672.

24. Cruveiller, S., Le Saux, J., Vallenet, D., Lajus, A., Bocs, S. and Médigue, C. (2005). MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. Nucleic Acids Res. 33: W471-479.

25. Vallenet, D., Labbare, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C. and Medigue, C. (2006). MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res. 34: 53-65.

26. Bryson, K., Loux, V., Bossy, R., Nicolas, P., Chaillou, S., van de Guchte, M.S., Penaud, S., Maguin, E., Hoebeke, M., Bessieres, P. and Gibrat, J.F. (2006). AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res. 34: 3533-3545.

27. Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 4:44-57.

28. Enault, F., Suhre, K. and Claverie, J-M. (2005). Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics. 6:247.

29. Almeida, L.G.P., Paixão, R., Souza, R.C., da Costa, G.C. Barrientos, F.J.A., dos Santos, M.T., de Almeida, D.F., and Vasconcelos, A.R. (2004). A System for Automated Bacterial (genome) Integrated Annotation—SABIA. Bioinformatics. 20: 2832-2833.

30. McNeil, L.K., Reich, C., Aziz, R.K., Bartels, D., Cohoon, M., Disz,T., Edwards, R.A., Gerdes, S., Hwang, K., Kubal, M., Margaryan, G. R., Meyer, F., Mihalo, W., Olsen, G.J., Olson, R., Osterman, A., Paarmann, D., Paczian, T., Parrello, B., Pusch,

G.D., Rodionov, D.A., Shi, X., Vassieva, O., Vonstein, V., Zagnitko, O., Xia, F., Zinner, J., Overbeek, R. and Stevens, R. (2007). The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. Nucleic Acids Res. 35: D347 – D353.

31. Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H-Y., Cohoon, M., de Cre´cy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen,N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D. Rodionov, D.A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva1, O., Ye, Y., Zagnith, O. and Vonstein, V. (2005). The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. Nucleic Acids Res. 33: 5691-5702.
32. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. and Kanehisa, M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35: W182–W185.
33. Kasukawa, T., Furuno, M., Nikaido, I., Bono, H., Hume, D. A., Bult, C., Hill, D. P.,Baldarelli, R., Gough, J., Kanapin, A., Matsuda, H., Schriml, L. M., Hayashizaki, Y., Okazaki, Y., Quackenbush, J. (2003). Development and evaluation of an automated annotation pipeline and cDNA annotation system. Genome Res. 13: 1542-51.

34. Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M. and Karp. P.D. (2005). EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res. 33:D334-D337.

35. Karp, P.D., Keseler, I.M., Shearer, A., Latendresse, M., Krummenacker, M., Paley, S.M., Paulsen, I., Collado-Vides, J., Gama-Castro, S., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza-Spínola, M.I., Bonavides-Martinez, C. and Ingraham, J. (2007). Multidimensional annotation of the Escherichia coli K-12 genome. Nucleic Acids Res. 35: 7577-7590.

36. Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C. and Gama-Castro. S. (2002). The Ecocyc Database. Nucleic Acids Res. 30: 56-58.

37. Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O., and Eisenberg, D. (2004). Prolinks: a database of protein functional Linkage derived from co-evolution. Genome Biology. 5: R35.

38. Price, M.N., Huang, K, H., Alm, E.J. and Arkin, A.P. (2005). A novel method for accurate predictions in all sequenced prokaryotes. Nucleic Acids Res. 33: 880-89.

39. Faith, J.J., Hayete, B., Thaden, J. T., Mogno, I., Weirzbowski, J., Cottarel, G., Kasif, S., Collins J.J. and Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biology. 5, E8.

40. Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H-R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L.L. and Bateman (2008). The Pfam Protein Families Database. Nucleic Acids Res. 36: 281-288.

41. Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. and Bairoch, A. (2004). Recent improvements to the PROSITE Database. Nucleic Acids Res. 32: D134-137.

42. Quevillon, E., Silventoinen, V., Pillai, S., Mulder, N., Apweiler, R. and Lopez, R.. (2005). InterProScan: proteins domains identifier. Nucleic Acids Res. 33: W116-W120.

43. Lima, T., Auchincloss, A.H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., Phan, I., Bougueleret, L. and Bairoch, A. (2009). HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Res. 37: D471-478.

44. Selengut, J.D., Haft, D.H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W.C., Richer, A.R. and White, O. (2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res. 35: D260-264.

45. Andrade, M.A., Brown, N.P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C. and Sander, C. (1999). Automated genome sequence analysis and annotation. Bioinformatics, 15: 391–412.

46. Frishman, D., Mokrejs, M., Kosykh, D., Kastenmüller, G., Kolesov, G., Zubrzycki, I., Gruber, C., Geier, B., Kaps, A., Albermann, K., Volz, A., Wagner, C., Fellenberg, M., Heumann, K. and Mewes, H.W. (2003). The PEDANT genome database. Nucleic Acids Res. 31: 207–211.

47. Harris, N. L. (1997). Genotator: a Workbench for Sequence Annotation. Genome Res.7:754–762.

48. Gaasterland, T. and Sensen, C.W. (1996). MAGPIE: automated genome interpretation. Trends Genet. 12: 76–88.

49. Gordon, P. M. K., Stromer, J., Turinsky, A.L., Xu, E., and Sensen, C.W. (1999). Bluejay: A Biological Sequence Browser featuring knowledge integration. Proceedings of the 13th Annual International Symposium on High Performance Computing Systems and Applications, Kingston, ON, Canada, pp. 183–194.

50. Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. and Puhler, A. (2003). GenDB - an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. 31: 2187–2195.

51. Van Domselaar, G.H., Stothard, P., Shrivastava, S., Cruz, J.A., Guo, A., Dong, X., Lu, P., Szafron, D., Greiner, R., and Wishart, D.S. (2005). BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res. 33: W455-W459.

52. Cannon, E., Magrane, M., Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., Donovan, C.O.and Apweil, R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Unitprot with Gene Ontology. Nucleic Acids Res. 32: D262-266.

53. Sundararaj, S., Guo, A., Habibi-Nazhad, B., Rouani, M., Stothard, P., Ellison, M. and Wishart, D.S. (2003). The Cyber Cell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli. NucleicAcids Res. 32: D293-D295.

54. The Gene Ontology Consortium. (2001). Creating the Gene Ontology Resource: Design and Implementation. Genome Res. 11: 1425–1433.

55. Cohen, K. B. and Hunter, L. (2008). Getting Started in Text Mining. PloS Computational Biology 4: 1-3.

56. Miotto, O., Tan, T.W. and Brusic, V. (2005). Supporting the curation of Biological databases with Reusable Text Mining. Genome Informatics 16: 32-44.

57. de Bruijin, B. and Martin, J. (2002). Getting to the (c)ore of knowledge: mining biomedical literature. Int J Med Inform. 67: 7-18.

Reference URL Links

U#	Tool Name	URL
U1	EcoCyc	http://www.ecocyc.org/
U2	BASys	http://wishart.biology.ualberta.ca/basys/cgi/gallery.pl
U3	BLASTN	ftp://ftp.ncbi.nlm.nih.gov/blast
U4	SSAHA	http://www.sanger.ac.uk/Software/analysis/SSAHA/
U5	GENSCAN	http://genes.mit.edu/GENSCAN.html
U 6	Genie	http://www.fruitfly.org/seq_tools/genie.html
U7	GeneMarkHMM	http://opal.biology.gatech.edu/GeneMark/
U8	Grail	http://compbio.ornl.gov/Grail-1.3/
U9	HMMGene	http://www.cbs.dtu.dk/services/HMMgene/
U10	BLASTX	ftp://ftp.ncbi.nlm.nih.gov/blast
U11	ePCR	http://www.ncbi.nlm.nih.gov/projects/e-pcr/
U12	SWISS-PROT	http://www.expasy.org/sprot/
U13	BLAT	http://genome.brc.mcw.edu/cgi-bin/hgBlat
U14	RepeatMasker	http://www.repeatmasker.org/
U15	COG	http://www.ncbi.nlm.nih.gov/COG/
U16	BLASTP	ftp://ftp.ncbi.nlm.nih.gov/blast
U17	TrEMBL	http://www.ebi.ac.uk/swissprot/
U18	PFAM	http://pfam.sanger.ac.uk/
U19	HMMER	http://www.psc.edu/general/software/packages/hmmer/
U20	PRINTS	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php
U21	MZEF	http://www.cshl.edu/OTT/html/mzef.html
U22	PROSITE	http://www.expasy.ch/prosite/
U23	ProDom	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/jj/prodomsrchjj.html
U24	PSI-BLAST	ftp://ftp.ncbi.nlm.nih.gov/blast
U25	SMART	http://smart.embl-heidelberg.de/
U26	BLOCKS	http://www.blocks.fhcrc.org/
U27	InterPro	http://www.ebi.ac.uk/proteome/
U28	FlyBase	http://flybase.bio.indiana.edu/
U29	Gene Ontology	http://www.geneontology.org/
U30	MouseGenome Database	http://www.informatics.jax.org/
U31	Gene Function Predictor	http://www.igs.cnrs-mrs.fr/phydbac/indexPS.html
U32	AGMIAL	http://genome.jouy.inra.fr/demo-agmial
U33	SABIA	http://www.phrap.org
U34	MaGe	http://www.genoscope.cns.fr/agc/mage
U35	MICHECK	http://www.genoscope.cns.fr/agc/tools/micheck
U36	DAVID	http://david.abcc.ncifcrf.gov/
U37	CCDB	http://redpoll.pharmacy.ualberta.ca/CCDB
U38	NMPDR	http://www.nmpdr.org
U39	KAAS	www.genome.jp/kegg/kaas/
U40	FASTA	http://www.ebi.ac.uk/fasta33/
U41	KEGG	ftp://ftp.genome.jp/pub/kegg/