Automated System for Gene Annotation and Metabolic Pathway Reconstruction Using General Sequence Databases

by **João M. P. Alves** and **Gregory A. Buck***

Center for the Study of Biological Complexity, Virginia Commonwealth University, P. O. Box 842030, Richmond, VA, 23284, USA
(phone: 1-804-828-2318; fax: 1-804-828-1397; e-mail: gabuck@vcu.edu)

Despite the growing number of genomes published or currently being sequenced, there is a relative paucity of software for functional classification of newly discovered genes and their assignment to metabolic pathways. Available software for such analyses has a very steep learning curve and requires the installation, configuration, and maintenance of large amounts of complex infrastructure, including complementary software and databases. Many such tools are restricted to one or a few data sources and classification schemes. In this work, we report an automated system for gene annotation and metabolic pathway reconstruction (ASGARD), which was designed to be powerful and generalizable, yet simple for the biologist to install and run on centralized, commonly available computers. It avoids the requirement for complex resources such as relational databases and web servers, as well as the need for administrator access to the operating system. Our methodology contributes to a more rapid investigation of the potential biochemical capabilities of genes and genomes by the biological researcher, and is useful in biochemical as well as comparative and evolutionary studies of pathways and networks.

Introduction. – The increasing availability of inexpensive and extremely efficient high-throughput DNA-sequencing methodologies [1–3] is driving the development of a large number of laboratories with the potential to sequence complete small-to-medium size genomes. The extensive sequencing of chromosomes and cDNA libraries from the larger genomes of more complex organisms is also now feasible for researchers outside of the large sequencing centers, as these second-generation sequencing technologies increase in availability and decrease in cost. Therefore, there is growing demand for analytical tools for sequence annotation and metabolic reconstruction to facilitate the interpretation of the huge amounts of data being generated.

Most of the software currently available for metabolic reconstruction requires that sequences be previously annotated in some form of controlled vocabulary, *e.g.*, Enzyme Commission (EC) [4] or Gene Ontology (GO) [5] numbers. Examples of such useful but limited projects include ArrayXPath [6], PathProcessor [7], and PathMAPA [8]. Other software pipelines can be very good at annotation, but offer no pathway identification capabilities and can be daunting to install and use by the biologist, *e.g.*, the very sophisticated Ensembl pipeline [9]. One interesting available tool which combines annotation and metabolic reconstruction in one package is KOBAS [10]. This tool is based on the 'Kyoto Encyclopedia of Genes and Genomes' (KEGG) orthology (KO) numbers [11] to assign function to new sequences and place them in known pathways. While a very useful and welcome effort with potential for great

contribution to systems biology approaches, KOBAS, as recognized by the authors [12], is implemented in a rather complex way that involves XML, relational databases, and many software dependencies, which add a significant challenge for installation and use even for the average computer-savvy biologist. While it would be ideal that each and every laboratory employed a bioinformaticist for such tasks, many if not most labs do not have the necessary resources.

Although KOBAS can currently be used on the original KOBAS servers through a Web browser, there are limitations on the number of sequences that can be analyzed in each run through this access. Moreover, there is a significant potential for heavy usage of these computers by investigators using remote access, leading to slow processing, and, like other Web-based resources, a significant risk that the site will not be permanently maintained and available. An even more significant concern is the exclusive dependency of KOBAS on KEGG data and KO numbers. While KEGG is becoming more comprehensive as the number of genomes included grows, there are still numerous organisms whose genomes have not been and may never be completely sequenced, and, therefore, will not be included in KEGG and similar metabolic databases which currently use only completely sequenced genomes. However, annotations and functional classifications of sequences from these organisms are frequently available in other, general databases such as Swiss-Prot or GenBank. These annotations are usually presented in the form of mappings of database sequence identifiers to functional classifications, and are provided by multiple research centers. For metabolic reconstructions, significant examples are the mappings of Swiss-Prot sequence identifiers to functional classifications like EC or GO (provided by the EMBL-EBI). Also important are mappings between different types of functional classification schemes; *e.g.*, mappings of GO to EC numbers (provided by the Gene Ontology Consortium). Such comprehensive resources permit the use of the full potential of general sequence databases and their associated resources in the reconstruction of metabolic pathways for newly sequenced organisms, without restricting the searches to a single source of information.

Herein, we describe ASGARD, a general, automated tool for sequence annotation and metabolic reconstruction using general databases. The input sequences can be either long DNA sequences (*e.g.*, contigs, chromosomes, complete genomes), or sets of genes or proteins. This tool was designed and documented to be accessible to as wide an audience of biological researchers as possible, with ease of installation and use as essential features. ASGARD is also designed to use plain text files of simple structure for data input, which makes the data easily extensible by the user, permitting integration of functional and pathway information from any available source. We have used ASGARD to successfully annotate and analyze the metabolic pathways in two genomes recently sequenced by our group, as well as genomes sequenced by other groups, confirming that ASGARD can significantly contribute to the complex and laborious tasks required by genomic biology.

Results and Discussion. – Our methodology was successfully applied to the two recently sequenced genomes of the apicomplexan protozoan *Cryptosporidium hominis* [13] and the oral bacterium *Streptococcus sanguinis* [14]. The proteins for the two genomes were analyzed and classified by ASGARD before publication of the

sequences and annotations in generally available databases. Examples of the output generated are presented in *Fig. 1* for the text format and *Fig. 2* for the graphical format. This discussion will focus on the results obtained for the bacterial proteins; *i.e.*, *S. sanguinis*, and perform a broad comparison of the performances of ASGARD and KOBAS on this dataset. We have also routinely used ASGARD for comparative metabolic and functional analyses (data not shown) of genomes from other organisms of interest, *e.g.*, *Cryptosporidium parvum* [15], *Plasmodium falciparum* [16], trypanosomatid protozoa [17–19], and a few streptococci [20][21].

a)

```
#Generated Thu Dec 21 00:14:38 2006 by ASGARD 0.9.7a, (September 26, 2006 by J)
#Infile used: ssang_pep_final.fasta.path_rec - This file: ssang_pep_final.fasta.paths
#Command: /home/jmalves/bin/asgard -w -i ssang_pep_final.fasta -p blastp -d /home/jmalve
#EC number,      Enzyme name
>Glycolysis / Gluconeogenesis 00010
1.1.1.27,       L-lactate dehydrogenase. ,      FOUND
2.7.1.11,       6-phosphofructokinase. , FOUND
2.7.1.69,       Protein-N(pi)-phosphohistidine--sugar phosphotransferase. ,      FOUND
2.7.2.-, Phosphotransferases with a carboxyl group as acceptor., FOUND
2.7.2.3, Phosphoglycerate kinase. ,      FOUND
4.1.2.13,       Fructose-bisphosphate aldolase. ,       FOUND
4.2.1.11        Phosphopyruvate hydratase.      FOUND
```

b)

```
>Glycolysis / Gluconeogenesis 00010
1.1.1.1, Alcohol dehydrogenase. , NF
1.1.1.2, Alcohol dehydrogenase (NADP(+)). ,     NF
1.1.1.27,       L-lactate dehydrogenase. ,      FOUND,  SSA_1221 Q3K1C8 4e-164 SSA_1221 Q5M3T6 3e-158 SSA_1221 Q60
1.1.1.71,       Alcohol dehydrogenase (NAD(P)(+)). ,    NF
1.1.99.8,       Alcohol dehydrogenase (acceptor). ,     NF
1.2.1.12,       Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating). ,   NF
1.2.1.3, Aldehyde dehydrogenase (NAD(+)). ,     NF
1.2.1.5, Aldehyde dehydrogenase (NAD(P)(+)). ,  NF
```

c)

```
SSA_0145,       Q8DTI9, GO:0003677 GO:0006350 GO:0006355,       2e-23
SSA_0171,       Q1CP28, GO:0043565,     2e-11
SSA_0171,       Q2ZZV3, GO:0043565,     7e-08
SSA_0173,       Q04I71, GO:0008168 GO:0008989 GO:0016740,       1e-95
SSA_0173,       Q97NE0, GO:0008168 GO:0008989 GO:0016740,       9e-95
SSA_0173,       Q3D7Q5, GO:0008168 GO:0016740,  7e-77
SSA_0173,       Q3D110, GO:0008168 GO:0016740,  9e-77
SSA_0173,       Q3DKG6, GO:0008168 GO:0016740,  1e-76
SSA_0173,       Q899Y6, GO:0008168 GO:0016740,  4e-76
SSA_0173,       Q300Y0, GO:0008168 GO:0008989 GO:0016740,       9e-69
SSA_0173,       Q5LXX9, GO:0008168 GO:0016740,  1e-63
```

Fig. 1. *Representative samples of the three output text files generated by ASGARD.* All three files start with an informative header (shown here only for the first file in *a*) showing the version of program used, the date the analysis was run, the files involved, identity of columns of data, and the full command issued. Every line of the header begins with the '#' symbol. The pathway summary file (*a*) contains the pathway titles and KEGG pathway numbers (lines starting with the '>' symbol), followed by three tab-delimited columns containing the EC or GO number, textual description of the function (when available), and a place holder keyword (FOUND). The detailed pathway report (*b*) differs from the summary in that it also reports the pathways components not found (NF) and has one additional column containing the identifier of the input sequence, its match, and the BLAST E-value of the match for all input sequences annotated with the function from that line. Finally, the full annotation file (*c*) contains the input sequence identifiers, match identifiers, functional classification, and *E*-value of the match, without sorting by pathway.

The analysis of *S. sanguinis* proteins identified 156 pathways, as defined by KEGG. In total, ASGARD identified 206 different EC numbers, assigned to 286 genes.

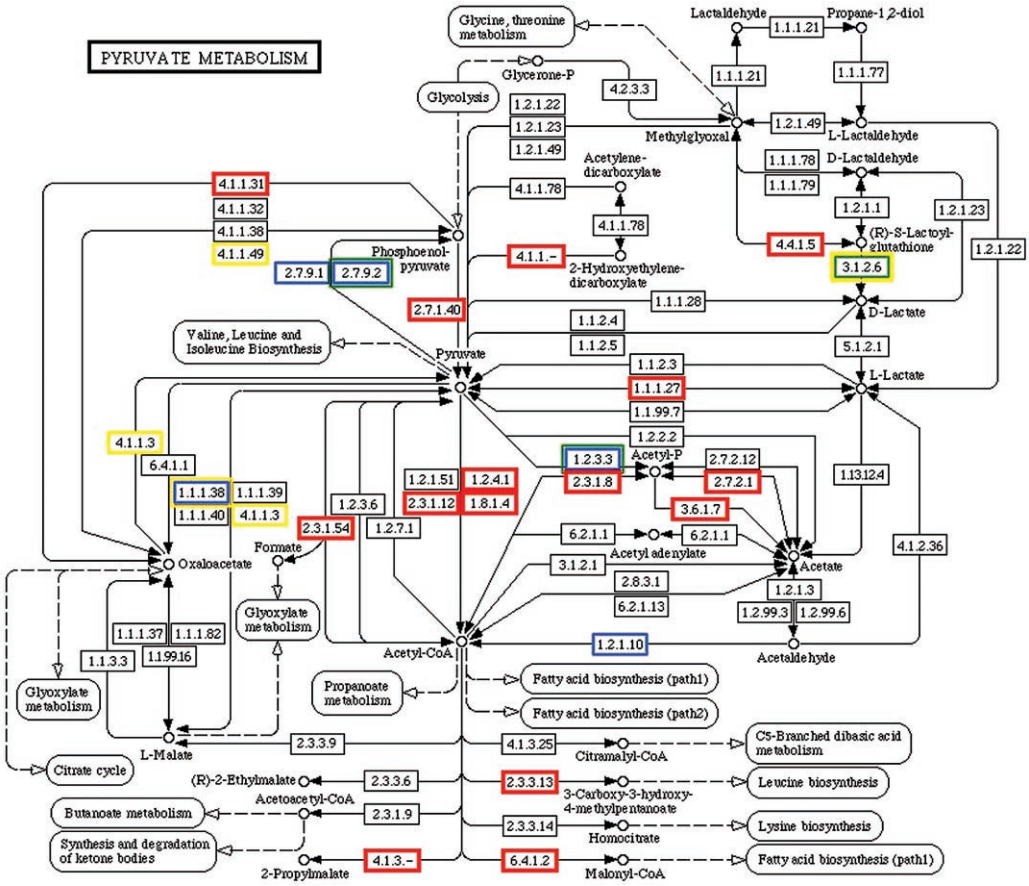

Fig. 2. *Sample graphical representation of the pyruvate metabolism pathway for three organisms.* Each organism is represented by one color (yellow, blue, or green). Enzymes surrounded by a double box are present in two of the three organisms, while enzymes surrounded by a thicker red box are present in all three organisms represented.

Furthermore, 1,369 other genes were annotated with GO classifications. Therefore, from the total 2,274 *S. sanguinis* protein-coding genes, 1,655 (72.78%) were assigned either an EC or GO number by ASGARD. Comparatively, analyzing this same genome with KOBAS identified 157 pathways, but annotated 1,344 of the genes analyzed (59.10%) with KO numbers. Comparison of the pathways identified shows 112 pathways in common. Further examination of the differences in the lists of identified pathways revealed that most of those identified exclusively by one of the programs were either not KEGG pathways used by KOBAS (categories of KEGG orthology not present as pathways, *e.g.*, 'other enzymes', 'other transporters', *etc.*) or were irrelevant to the bacterial system, *e.g.*, human disease pathways, eukaryotic signaling systems, *etc.* When these obvious spurious identifications were removed, 14 pathways remained which were only identified by KOBAS, and 15 which were identified only by

ASGARD. Close inspection of these remaining pathways showed that all but two ('protein export' and 'ribosome', identified by KOBAS) have no significance and are clearly not functional units, since these pathways contained only one or two or a few scattered enzymes, generally also needed for other, more complete pathways. KOBAS performed very well in the identification of the proteins from the ribosome and protein-export pathways, finding 52 ribosomal proteins, as well as the components for protein-export previously manually identified, namely the TatC protein, signal-recognition particle components and the Sec-dependent pathway proteins [14]. The only error in the protein-export pathway was the assignment of a cytoplasmic Mg-dependent DNase (TatD) to this pathway. We believe this error was due to a misclassification by KEGG and not to an error in processing by KOBAS, since the KO number for this protein (K03424, Mg-dependent DNase) is indeed listed as belonging to the protein-export pathway (ko03060). The error seems to be based on obsolete theories about the function of this protein [22].

It is significant to note that the two pathways identified only by KOBAS were actually not pathways usually associated with metabolism, but more structural components of the cell: 'protein export' and 'ribosome'. These categories are comprised of several proteins which are classified only with KO numbers within the KEGG system, but no EC or GO numbers. It was expected that these types of pathways would not be identified by ASGARD, as the software currently only uses EC and GO numbers. However, the flexible nature of ASGARD permits incorporation of any classification scheme or sequence database, in either addition to or substituted for those currently employed. Thus, it would be possible to add KO numbers and the KEGG sequence databases to the data used by ASGARD, and use it together with the default EC and GO numbers and UniProt100 database. To do so, a mapping between the KEGG sequence database identifiers and the KO numbers is necessary. Such a mapping can be readily obtained by joining all KO list files for all genomes present in KEGG, or, in a less practical and more error-prone way, by parsing the information from the FASTA header from the available file containing all sequences.

As an illustrative example of these flexible extension capabilities of ASGARD, we complemented the standard data with EC and GO numbers, protein sequences, and corresponding mappings of other available streptococcal genomes at the time of analysis of the *S. sanguinis* genome [14]. We did this because we observed that mappings available for the general databases had a relatively poor representation of annotated streptococcal genomes. Thus, we gathered the relevant data from GenBank annotation and KEGG files for the streptococcal genomes, parsed the EC and GO numbers and sequence identifiers, and converted them to the format required for ASGARD. This complementation significantly enhanced the annotation of *S. sanguinis* (data not shown), greatly reducing the need for manual annotation and metabolic reconstruction.

Future developments of the ASGARD system might include the creation of accessory programs and documentation for the conversion and incorporation of data from the most widely used pathway databases (KEGG, Reactome, BioCyc, *etc.*) to ASGARD, to provide the final user with more choice of data sources when running the program. Interestingly, a similar complementation of the data was not necessary for the analysis of the *C. hominis* genome [13], indicating an eukaryotic bias in the standard

data used and highlighting the importance of being able to include additional data sources in the analysis, depending on the specific nature of the project.

Conclusions. – We have shown ASGARD to be an efficient, easy to install and use program, yielding very good results in the study and comparison of genomes. The program is also easily extensible: any source of functional information or sequence database can be incorporated or substituted for the ones originally used in this work, eliminating reliance on a single data source and increasing the potential use of the program for very different types of projects. While we believe no automatic, computerized solution can substitute for expert manual annotation, ASGARD can significantly expedite analysis of the massive amounts of data being generated by the biological community.

We would like to thank the other members of our research laboratory for thoughtful and constructive input and comments, and *Yue Zhao* for help with the graphical aspects of the software. This work was supported by USPHS grants AI50425, AI50196, and AI046418 from the *National Institute of Allergy and Infectious Disease*, and DE12882 from the *National Institute of Dental and Craniofacial Research*.

Experimental Part

Required Data. ASGARD uses annotation data derived from a variety of sources to assign sequences to biological pathways. Among the available general databases, Swiss-Prot-derived databases; *e.g.*, UniProt [23], generally have the most complete annotation. The main advantage to using these databases is their broad taxonomic coverage, without restriction to organisms for which complete genomes are available. Previous attempts at metabolic reconstruction from genomic data using similarity; *e.g.*, KOBAS, have relied exclusively on smaller databases like KEGG, which only contain data from completely sequenced genomes. ASGARD does not rely on the KEGG database of sequences for the annotation. However, since KEGG is a very comprehensive and well-designed metabolic database for which the pathways can be automatically imaged quite easily, ASGARD does use KEGG as the source for pathway definitions (list of enzymes for each pathway) and graphical representations of pathways. Although we have used KEGG as an information source, ASGARD is capable of using any pathway definition source for which the information can be converted into a simple two-column representation.

It is necessary to know the functional annotations of database sequences to be able to assign function to new sequences based on similarity. There are several mappings between Swiss-Prot identifiers and functional information; *e.g.*, GO and EC numbers. Classifications based on these functional mappings are most commonly used, as they constitute one of the bases for pathway databases like KEGG, Biocyc [24], and others. The crucial feature of these classification schemes is the use of a defined and stable vocabulary for functional description of biological sequences, permitting general use of annotations that would be impossible if this information were encoded only in natural human language. All data files required by ASGARD can be downloaded automatically (see below) using the accompanying *download_ASGARD_data* program. An overview of the data and its processing is presented in *Fig. 3*.

The database we generally use with ASGARD is UniRef100, a non-redundant reference cluster from UniProt (http://www.uniprot.org). This database was used both in FASTA format (*uniprot100.fasta*) and a BLAST-compatible format (*UniProt100.p**). Gene-Ontology definitions and mappings of GO to EC numbers were downloaded from the Gene Ontology Consortium site (http://www.geneontology.org/). Mappings of UniProt sequence identifiers with GO numbers are obtained from the EMBL-EBI (http://www.ebi.ac.uk) and converted to a Berkeley DB database file (*sp2go.db*). Definitions for EC numbers and categories (used for annotation of the pathways list file *kegg_pathways*) and mappings of UniProt to EC numbers (converted to database file *sp2ec.db*) were from the Swiss Institute of Bioinformatics' Enzyme Nomenclature database (http://www.expasy.org/). Finally, pathway names, definitions, and images were from KEGG (http://www.genome.jp/kegg/).

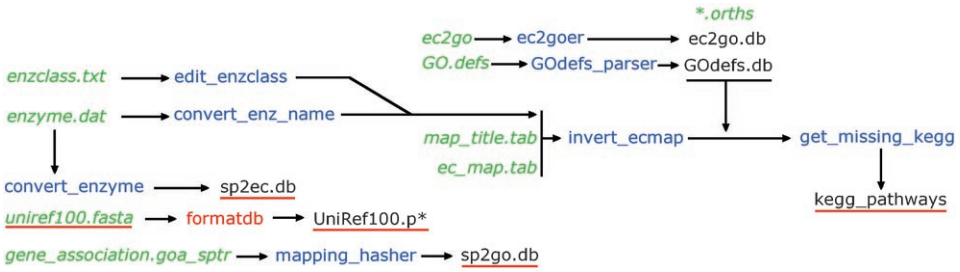

Fig. 3. *General scheme describing the conversion of data downloaded, as described in the text, to required formats using* format_ASGARD_data. Data files that will be used at runtime (*Fig. 4*) are underlined in red; downloaded data files are in green italic font; external software requirements are depicted in red font; conversion steps are in blue font (these are also some of the scripts that can be run independently in case of manual updating and/or complementation of the data by the user).

Software and Hardware Requirements. External software requirements for ASGARD include the NCBI BLAST [25] and the EMBOSS [26] packages. From the EMBOSS package, ASGARD employs *splitter*, which is required only when the analysis starts with long DNA sequences that must be split for further analysis (described below). Splitter is not required if the user intends to analyze gene or protein sequences. We redistribute *splitter* in compiled form for a few common POSIX-based systems to minimize the complexity of the installation procedures (EMBOSS is a very powerful suite of programs, and its full installation can be a demanding task in some cases).

For parallel operation in a high-performance computing environment, ASGARD depends on a batch queuing system; *e.g.*, the Sun Grid Engine v 5.3 (SGE, http://gridengine.sunsource.net) or the Portable Batch System (PBS, http://www.openpbs.org), both of which have been successfully used in conjunction with the program. More specifically, our software uses the *qsub* command to submit the multiple BLAST search jobs to available nodes. Since both scheduling systems use the same command name and syntax, there is no need to modify the code for ASGARD to work using either system.

ASGARD can also be used on single processor machines or on parallel machines without SGE or PBS. Thus, a user would launch ASGARD and subsequently manually run the resultant BLAST jobs in a separate terminal session, or even, if warranted, on separate computers. ASGARD will be in waiting mode, and automatically resume processing after all BLAST jobs are completed and present in the temporary directory specifically created for the current run (details of these operations are available in the program documentation). However, execution times on single processor machines will extend for several days when searching for a few thousand genes, or weeks when searching small genomes against UniProt. A solution is a hybrid environment in which BLAST searches would be performed on some machines, while the final processing by ASGARD is efficiently completed on a single unit. ASGARD has special modes included for easy implementation of this modular, multi-machine processing.

ASGARD is written in Perl, and tested using Perl versions 5.6.1 and 5.8.5. All modules required by ASGARD and other accompanying programs are usually included with currently standard Perl installations. A possible exception is the GD graphics Perl module, which is required for drawing pathway diagrams. The GD graphics module can be readily installed from the CPAN system (http://www.cpan.org). The operating system environments used for development and use were variants of *Linux*, although the program should also work on similar systems containing the standard command line tools and shell commands. ASGARD is free software distributed under the terms of the GNU General Public License version 2 (http://www.gnu.org/licenses/licenses.html).

Analysis of the data by KOBAS was performed on the web interface (http://kobas.cbi.pku.edu.cn/index.jsp). For comparison with ASGARD, we used the most similar parameters possible, considering the first top ten hits and using E-value < 1e-6 threshold for BLAST (the defaults for KOBAS are top five hits and E-value < 1e-5).

The Algorithm. The general schemes of ASGARD required data files and runtime operation are presented in *Figs. 3* and *4*. In brief, the program begins by making general checks for the presence of required data files (previously generated by the *format_ASGARD_data* program), input file, and required parameters. The input for ASGARD is a multi-sequence FASTA file containing either gene, protein, or long DNA sequences (*e.g.*, contigs or whole chromosomes). If the input is long DNA sequences, each sequence in the input file will be split by EMBOSS *splitter* program into pieces of a determined size (default 300 nucleotides). Gene or protein sequences are used directly, without fragmentation. These sequences are searched using BLAST (default E < 1e-6) against the UniRef100 database of proteins to find the top ten matches. ASGARD uses mappings of UniProt identifiers to functional classifications to identify the first sequence among these matches to possess an annotation and assigns this annotation to the user sequence. The program prioritizes EC numbers over GO numbers, and, therefore, will assign an EC number first if possible. If no EC number is present among the annotations for the best ten matches, a GO number will be assigned. If neither an EC number nor a GO number is identified, the sequence cannot be annotated by ASGARD.

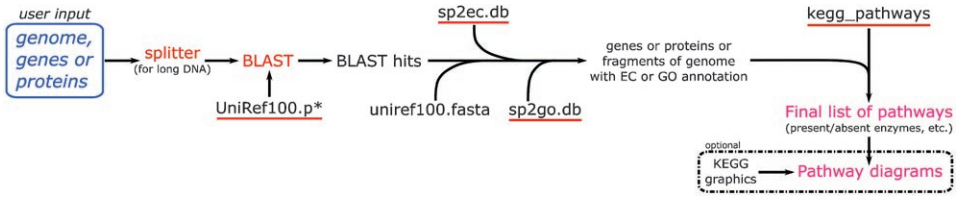

Fig. 4. *Requirements and general scheme describing the operation of ASGARD.* Data files previously formatted by *format_ASGARD_data* are underlined in red (*Fig. 3*); external software requirements are depicted in red font; the final outputs are in magenta. Splitter, sp2ec.db, sp2go.db, UniRef100.p*, and UniRef100.fasta are described in the text.

Finally, ASGARD screens the pathway definition list for each component of each pathway. The output consists of three files: *1*) a summary of the identified sequences and pathways (a file with *.paths* extension); *2*) a file containing the details of the pathway components (whether found or not, and, if found, the gene identifiers, top match identifiers and E-value of the matches, *.paths_detail* extension); and *3*) a file with the matches, annotation and BLAST E-values of all input sequences that matched the database and had an annotation, regardless of whether they belong to one of the pathways in the definition list (*.path_rec* extension). An example of each type of file is presented in *Fig. 1*. Graphical representations of the reconstructed pathways (*Fig. 2*) can also be generated after ASGARD is run, by using the pathway summary file as input to the included *color_map* program. *Color_map* uses KEGG pathway images and coordinates to draw colored borders around the components identified by ASGARD. The program will compare up to three different datasets from different input sequences. Different colors (which can be changed by the user) are displayed depending on which datasets contain the genes. Thus, two-color boxes are created when enzymes are identified in two of three data sets, and one thicker red box when an enzyme is identified in three datasets. Thus, *color_map* can be quite useful for comparative studies of metabolism of two or three organisms. Four colors are sufficient to represent all combinations of presence of components for up to three sets of sequences.

REFERENCES

[1] R. F. Service, *Science* **2006**, *311*, 1544.

[2] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu,

V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, J. M. Rothberg, *Nature* **2005**, *437*, 376.

[3] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, G. M. Church, *Science* **2005**, *309*, 1728.

[4] A. Bairoch, *Nucleic Acids Res.* **2000**, *28*, 304.

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, *Nat. Genet.* **2000**, *25*, 25.

[6] H. J. Chung, M. Kim, C. H. Park, J. Kim, J. H. Kim, *Nucleic Acids Res.* **2004**, *32*, W460.

[7] P. Grosu, J. P. Townsend, D. L. Hartl, D. Cavalieri, *Genome Res.* **2002**, *12*, 1121.

[8] D. Pan, N. Sun, K. H. Cheung, Z. Guan, L. Ma, M. Holford, X. Deng, H. Zhao, *BMC Bioinformatics* **2003**, *4*, 56.

[9] V. Curwen, E. Eyras, T. D. Andrews, L. Clarke, E. Mongin, S. M. J. Searle, M. Clamp, *Genome Res.* **2004**, *14*, 942.

[10] X. Mao, T. Cai, J. G. Olyarchuk, L. Wei, *Bioinformatics* **2005**, *21*, 3787.

[11] M. Kanehisa, S. Goto, *Nucleic Acids Res.* **2000**, *28*, 27.

[12] J. Wu, X. Mao, T. Cai, J. Luo, L. Wei, *Nucleic Acids Res.* **2006**, *34*, W720.

[13] P. Xu, G. Widmer, Y. Wang, L. S. Ozaki, J. M. Alves, M. G. Serrano, D. Puiu, P. Manque, D. Akiyoshi, A. J. Mackey, W. R. Pearson, P. H. Dear, A. T. Bankier, D. L. Peterson, M. S. Abrahamsen, V. Kapur, S. Tzipori, G. A. Buck, *Nature* **2004**, *431*, 1107.

[14] P. Xu, J. M. Alves, T. Kitten, A. Brown, Z. Chen, L. S. Ozaki, P. Manque, X. Ge, M. G. Serrano, D. Puiu, S. Hendricks, Y. Wang, M. D. Chaplin, D. Akan, S. Paik, D. L. Peterson, F. L. Macrina, G. A. Buck, *J. Bacteriol.* **2007**, *189*, 3166.

[15] M. S. Abrahamsen, T. J. Templeton, S. Enomoto, J. E. Abrahante, G. Zhu, C. A. Lancto, M. Deng, C. Liu, G. Widmer, S. Tzipori, G. A. Buck, P. Xu, A. T. Bankier, P. H. Dear, B. A. Konfortov, H. F. Spriggs, L. Iyer, V. Anantharaman, L. Aravind, V. Kapur, *Science* **2004**, *304*, 441.

[16] M. J. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, B. Barrell, *Nature* **2002**, *419*, 498.

[17] N. M. El-Sayed, P. J. Myler, D. C. Bartholomeu, D. Nilsson, G. Aggarwal, A. N. Tran, E. Ghedin, E. A. Worthey, A. L. Delcher, G. Blandin, S. J. Westenberger, E. Caler, G. C. Cerqueira, C. Branche, B. Haas, A. Anupama, E. Arner, L. Aslund, P. Attipoe, E. Bontempi, F. Bringaud, P. Burton, E. Cadag, D. A. Campbell, M. Carrington, J. Crabtree, H. Darban, J. F. da Silveira, P. de Jong, P. K. Edwards, P. T. Englund, G. Fazelina, T. Feldblyum, M. Ferella, A. C. Frasch, K. Gull, D. Horn, L. Hou, Y. Huang, E. Kindlund, M. Klingbeil, S. Kluge, H. Koo, D. Lacerda, M. J. Levin, H. Lorenzi, T. Louie, C. R. Machado, R. McCulloch, A. McKenna, Y. Mizuno, J. C. Mottram, S. Nelson, S. Ochaya, K. Osoegawa, G. Pai, M. Parsons, M. Pentony, U. Pettersson, M. Pop, J. L. Ramirez, J. Rinta, L. Robertson, S. L. Salzberg, D. O. Sanchez, A. Seyler, R. Sharma, J. Shetty, A. J. Simpson, E. Sisk, M. T. Tammi, R. Tarleton, S. Teixeira, S. Van Aken, C. Vogt, P. N. Ward, B. Wickstead, J. Wortman, O. White, C. M. Fraser, K. D. Stuart, B. Andersson, *Science* **2005**, *309*, 409.

[18] M. Berriman, E. Ghedin, C. Hertz-Fowler, G. Blandin, H. Renauld, D. C. Bartholomeu, N. J. Lennard, E. Caler, N. E. Hamlin, B. Haas, U. Bohme, L. Hannick, M. A. Aslett, J. Shallom, L. Marcello, L. Hou, B. Wickstead, U. C. Alsmark, C. Arrowsmith, R. J. Atkin, A. J. Barron, F. Bringaud, K. Brooks, M. Carrington, I. Cherevach, T. J. Chillingworth, C. Churcher, L. N. Clark, C. H. Corton, A. Cronin, R. M. Davies, J. Doggett, A. Djikeng, T. Feldblyum, M. C. Field, A. Fraser, I. Goodhead, Z. Hance, D. Harper, B. R. Harris, H. Hauser, J. Hostetler, A. Ivens, K. Jagels, D. Johnson, J. Johnson, K. Jones, A. X. Kerhornou, H. Koo, N. Larke, S. Landfear, C. Larkin, V. Leech,

A. Line, A. Lord, A. Macleod, P. J. Mooney, S. Moule, D. M. Martin, G. W. Morgan, K. Mungall, H. Norbertczak, D. Ormond, G. Pai, C. S. Peacock, J. Peterson, M. A. Quail, E. Rabbinowitsch, M. A. Rajandream, C. Reitter, S. L. Salzberg, M. Sanders, S. Schobel, S. Sharp, M. Simmonds, A. J. Simpson, L. Tallon, C. M. Turner, A. Tait, A. R. Tivey, S. Van Aken, D. Walker, D. Wanless, S. Wang, B. White, O. White, S. Whitehead, J. Woodward, J. Wortman, M. D. Adams, T. M. Embley, K. Gull, E. Ullu, J. D. Barry, A. H. Fairlamb, F. Opperdoes, B. G. Barrell, J. E. Donelson, N. Hall, C. M. Fraser, S. E. Melville, N. M. El-Sayed, *Science* **2005**, *309*, 416.

[19] A. C. Ivens, C. S. Peacock, E. A. Worthey, L. Murphy, G. Aggarwal, M. Berriman, E. Sisk, M. A. Rajandream, E. Adlem, R. Aert, A. Anupama, Z. Apostolou, P. Attipoe, N. Bason, C. Bauser, A. Beck, S. M. Beverley, G. Bianchettin, K. Borzym, G. Bothe, C. V. Bruschi, M. Collins, E. Cadag, L. Ciarloni, C. Clayton, R. M. Coulson, A. Cronin, A. K. Cruz, R. M. Davies, J. De Gaudenzi, D. E. Dobson, A. Duesterhoeft, G. Fazelina, N. Fosker, A. C. Frasch, A. Fraser, M. Fuchs, C. Gabel, A. Goble, A. Goffeau, D. Harris, C. Hertz-Fowler, H. Hilbert, D. Horn, Y. Huang, S. Klages, A. Knights, M. Kube, N. Larke, L. Litvin, A. Lord, T. Louie, M. Marra, D. Masuy, K. Matthews, S. Michaeli, J. C. Mottram, S. Muller-Auer, H. Munden, S. Nelson, H. Norbertczak, K. Oliver, S. O'Neil, M. Pentony, T. M. Pohl, C. Price, B. Purnelle, M. A. Quail, E. Rabbinowitsch, R. Reinhardt, M. Rieger, J. Rinta, J. Robben, L. Robertson, J. C. Ruiz, S. Rutter, D. Saunders, M. Schafer, J. Schein, D. C. Schwartz, K. Seeger, A. Seyler, S. Sharp, H. Shin, D. Sivam, R. Squares, S. Squares, V. Tosato, C. Vogt, G. Volckaert, R. Wambutt, T. Warren, H. Wedler, J. Woodward, S. Zhou, W. Zimmermann, D. F. Smith, J. M. Blackwell, K. D. Stuart, B. Barrell, P. J. Myler, *Science* **2005**, *309*, 436.

[20] D. Ajdic, W. M. McShan, R. E. McLaughlin, G. Savic, J. Chang, M. B. Carson, C. Primeaux, R. Tian, S. Kenton, H. Jia, S. Lin, Y. Qian, S. Li, H. Zhu, F. Najar, H. Lai, J. White, B. A. Roe, J. J. Ferretti, *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14434.

[21] J. Hoskins, W. E. Alborn Jr., J. Arnold, L. C. Blaszczak, S. Burgett, B. S. DeHoff, S. T. Estrem, L. Fritz, D. J. Fu, W. Fuller, C. Geringer, R. Gilmour, J. S. Glass, H. Khoja, A. R. Kraft, R. E. Lagace, D. J. LeBlanc, L. N. Lee, E. J. Lefkowitz, J. Lu, P. Matsushima, S. M. McAhren, M. McHenney, K. McLeaster, C. W. Mundy, T. I. Nicas, F. H. Norris, M. O'Gara, R. B. Peery, G. T. Robertson, P. Rockey, P. M. Sun, M. E. Winkler, Y. Yang, M. Young-Bellido, G. Zhao, C. A. Zook, R. H. Baltz, S. R. Jaskunas, P. R. Rosteck Jr., P. L. Skatrud, J. I. Glass, *J. Bacteriol.* **2001**, *183*, 5709.

[22] M. Wexler, F. Sargent, R. L. Jack, N. R. Stanley, E. G. Bogsch, C. Robinson, B. C. Berks, T. Palmer, *J. Biol. Chem.* **2000**, *275*, 16717.

[23] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, L. S. Yeh, *Nucleic Acids Res.* **2004**, *32*, D115.

[24] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, L. Stein, *Nucleic Acids Res.* **2005**, *33*, D428.

[25] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **1997**, *25*, 3389.

[26] P. Rice, I. Longden, A. Bleasby, *Trends Genet.* **2000**, *16*, 276.

Received February 23, 2007