AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

METHODS AND PROTOCOLS
Novel Systems Biology Techniques

Curated BLAST for Genomes

Morgan N. Price,[a] Adam P. Arkin[a]

[a]Lawrence Berkeley National Laboratory, Berkeley, California, USA

ABSTRACT Curated BLAST for Genomes finds candidate genes for a process or an enzymatic activity within a genome of interest. In contrast to annotation tools, which usually predict a single activity for each protein, Curated BLAST asks if any of the proteins in the genome are similar to characterized proteins that are relevant. Given a query such as an enzyme's name or an EC number, Curated BLAST searches the curated descriptions of over 100,000 characterized proteins, and it compares the relevant characterized proteins to the predicted proteins in the genome of interest. In case of errors in the gene models, Curated BLAST also searches the six-frame translation of the genome. Curated BLAST is available at http://papers.genomics.lbl.gov/curated.

IMPORTANCE Given a microbe's genome sequence, we often want to predict what capabilities the organism has, such as which nutrients it requires or which energy sources it can use. Or, we know the organism has a capability and we want to find the genes involved. Scientists often use automated gene annotations to find relevant genes, but automated annotations are often vague or incorrect. Curated BLAST finds candidate genes for a capability without relying on automated annotations. First, Curated BLAST finds proteins (usually from other organisms) whose functions have been studied experimentally and whose curated descriptions match a query. Then, it searches the genome of interest for similar proteins and returns a list of candidates. Curated BLAST is fast and often finds relevant genes that are missed by automated annotation.

KEYWORDS annotation

Given the genome sequence for an organism of interest, we often want to know whether or not it encodes a certain capability and which proteins might be involved. To support this, many genomics websites support searching for proteins whose annotations match a text query. However, annotation tools will usually provide one predicted function for each protein, and these predictions are often incorrect (i.e., reference 1). So, searching through annotations may not be the best way to find proteins that are involved in a process.

Instead, we propose that given a text query, we can identify experimentally characterized proteins (usually from other organisms) that are relevant. Then, we can search in the genome of interest for proteins that are similar to these characterized proteins. For enzymes, this approach obviates the need to predict the substrate specificity (which is often not possible). Instead, we identify candidates that are similar to characterized proteins that have activities of interest.

We implemented this approach in a web-based tool called Curated BLAST for Genomes (http://papers.genomics.lbl.gov/curated). It relies on a collection of over 100,000 characterized proteins, and it usually takes just a few seconds per query.

RESULTS

Finding a missing annotation. For example, consider searching for "perchlorate" in *Azospira oryzae* PS (Fig. 1). It takes a few seconds for Curated BLAST to identify three

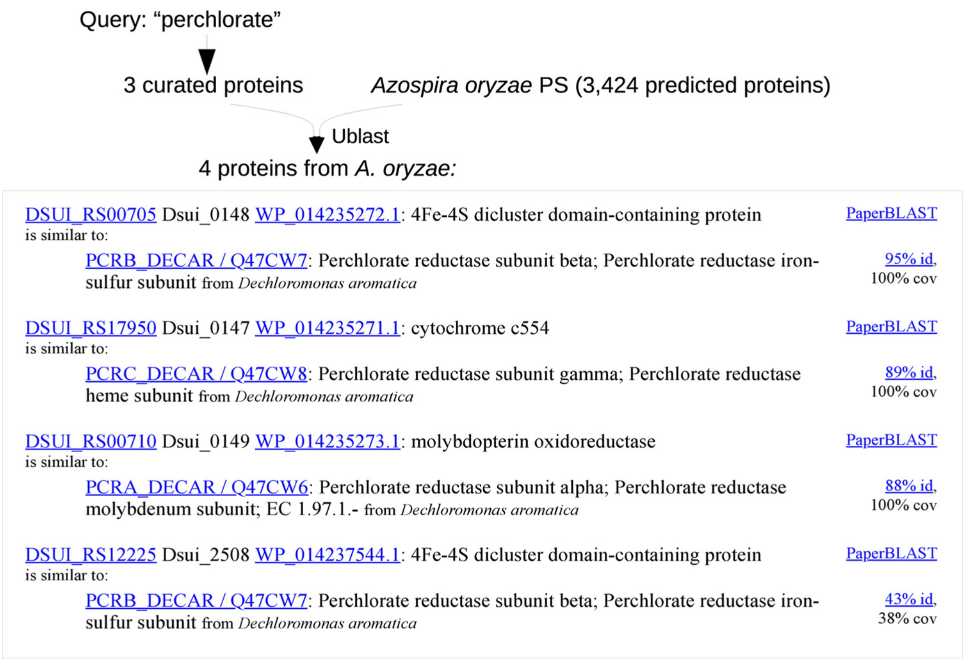
Citation Price MN, Arkin AP. 2019. Curated BLAST for genomes. mSystems 4:e00072-19. https://doi.org/10.1128/mSystems.00072-19.

Editor Casey S. Greene, University of Pennsylvania

Address correspondence to Morgan N. Price, morgannprice@yahoo.com, or Adam P. Arkin, aparkin@lbl.gov.

Curated BLAST for genomes: find characterized proteins that match a query, and then search a genome for similar proteins.

Received 1 February 2019
Accepted 13 March 2019
Published 26 March 2019


Query: "perchlorate"
3 curated proteins
Azospira oryzae PS (3,424 predicted proteins)
Ublast
4 proteins from *A. oryzae:*

DSUI_RS00705 Dsui_0148 WP_014235272.1: 4Fe-4S dicluster domain-containing protein — PaperBLAST
is similar to:
PCRB_DECAR / Q47CW7: Perchlorate reductase subunit beta; Perchlorate reductase iron-sulfur subunit from *Dechloromonas aromatica* — 95% id, 100% cov

DSUI_RS17950 Dsui_0147 WP_014235271.1: cytochrome c554 — PaperBLAST
is similar to:
PCRC_DECAR / Q47CW8: Perchlorate reductase subunit gamma; Perchlorate reductase heme subunit from *Dechloromonas aromatica* — 89% id, 100% cov

DSUI_RS00710 Dsui_0149 WP_014235273.1: molybdopterin oxidoreductase — PaperBLAST
is similar to:
PCRA_DECAR / Q47CW6: Perchlorate reductase subunit alpha; Perchlorate reductase molybdenum subunit; EC 1.97.1.- from *Dechloromonas aromatica* — 88% id, 100% cov

DSUI_RS12225 Dsui_2508 WP_014237544.1: 4Fe-4S dicluster domain-containing protein — PaperBLAST
is similar to:
PCRB_DECAR / Q47CW7: Perchlorate reductase subunit beta; Perchlorate reductase iron-sulfur subunit from *Dechloromonas aromatica* — 43% id, 38% cov

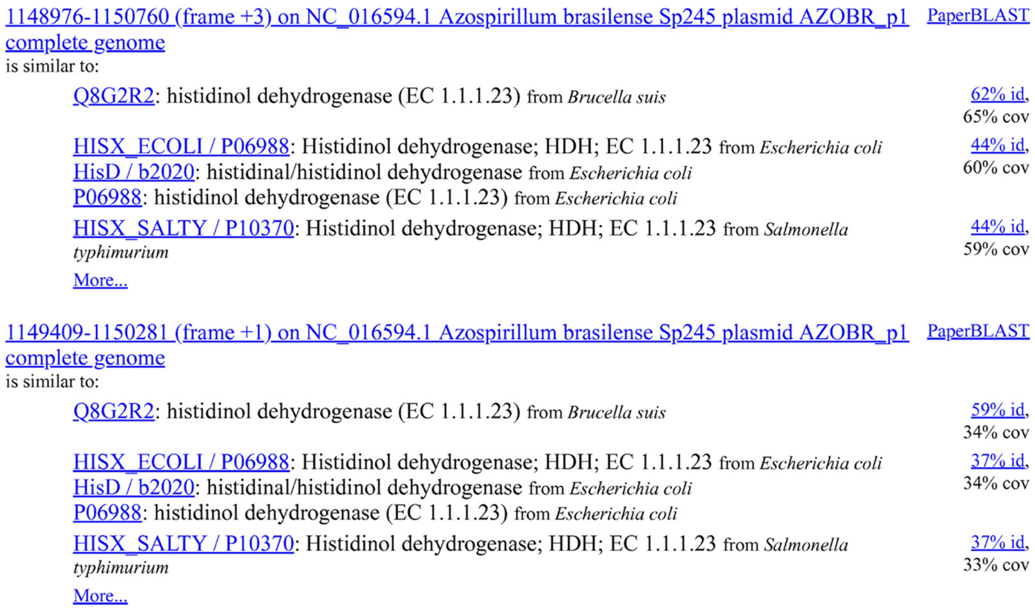
FIG 1 An example of Curated BLAST. We show an overview of the process and a screenshot from the website. id, percent identity of the amino acid sequence; cov, percent coverage of the characterized protein.

proteins in the genome that are over 80% identical, over their full length, to the three subunits of a putative perchlorate reductase. The putative perchlorate reductase is from a related species (*Azospira oryzae* used to be named *Dechlorosoma suillum*), was identified by genetic approaches (2), and is curated in Swiss-Prot (3). Two of the proteins from the PS strain have actually been demonstrated to reduce perchlorate (PcrAB) (4), but this is not reflected in any of the databases. It might still seem easy to annotate the proteins in the PS strain, given that they are so similar to proteins in Swiss-Prot, but as of November 2018, neither RefSeq (5), RAST (6), nor KEGG (7) annotates any of these proteins as perchlorate reductase. In RAST and KEGG, these proteins are misannotated as a nitrate reductase. Perchlorate reductase illustrates how it can be easier to find a protein with Curated BLAST than by using gene annotations.

Curated BLAST also found another protein in the genome of *A. oryzae*, Dsui_2508, that has some similarity to perchlorate reductase. The alignment covers less than 50% of the perchlorate reductase subunit ("38% cov"), and the similarity is modest ("43% id."), which suggests that Dsui_2508 might have a different function. To identify other potential functions for the proteins that are returned, the search results include a link to PaperBLAST (8), which finds papers and curated entries about homologs of a protein of interest. Clicking on the PaperBLAST link shows that Dsui_2508 is 57% identical, over 85% of its length, to a subunit of a tetrathionate reductase, so Dsui_2508 is not likely to be involved in perchlorate reduction.

As illustrated by this example, alignments with a high percent identity and a high percent coverage are the most likely to be relevant. So, Curated BLAST sorts the proteins that it finds in the genome of interest by the percent identity × percent coverage of their best alignment to a (relevant) curated protein. Similarly, for each predicted protein that has hits, Curated BLAST shows the curated proteins that align with the highest percent identity × percent coverage first.

Finding candidates even when the gene models are incorrect. Another challenge in finding proteins in a genome is that the protein of interest may be missing from the list of predicted proteins. Proteomics studies often identify proteins that were missed in the computational gene annotation (i.e., reference 9). Insertion or deletion



1148976-1150760 (frame +3) on NC_016594.1 Azospirillum brasilense Sp245 plasmid AZOBR_p1 complete genome PaperBLAST
is similar to:

Q8G2R2: histidinol dehydrogenase (EC 1.1.1.23) from *Brucella suis* 62% id, 65% cov

HISX_ECOLI / P06988: Histidinol dehydrogenase; HDH; EC 1.1.1.23 from *Escherichia coli* 44% id, 60% cov
HisD / b2020: histidinal/histidinol dehydrogenase from *Escherichia coli*
P06988: histidinol dehydrogenase (EC 1.1.1.23) from *Escherichia coli*

HISX_SALTY / P10370: Histidinol dehydrogenase; HDH; EC 1.1.1.23 from *Salmonella typhimurium* 44% id, 59% cov

More...

1149409-1150281 (frame +1) on NC_016594.1 Azospirillum brasilense Sp245 plasmid AZOBR_p1 complete genome PaperBLAST
is similar to:

Q8G2R2: histidinol dehydrogenase (EC 1.1.1.23) from *Brucella suis* 59% id, 34% cov

HISX_ECOLI / P06988: Histidinol dehydrogenase; HDH; EC 1.1.1.23 from *Escherichia coli* 37% id, 34% cov
HisD / b2020: histidinal/histidinol dehydrogenase from *Escherichia coli*
P06988: histidinol dehydrogenase (EC 1.1.1.23) from *Escherichia coli*

HISX_SALTY / P10370: Histidinol dehydrogenase; HDH; EC 1.1.1.23 from *Salmonella typhimurium* 37% id, 33% cov

More...

FIG 2 Six-frame search results for “histidinol dehydrogenase” against *Azospirillum brasilense* Sp245.

errors within the sequence of a protein coding gene can also prevent the protein from being identified. This may be a common problem with long-read single-molecule sequencing: many of the resulting genomes have a suspiciously large number of genes that are disrupted by frameshifts (Mick Watson, 8 March 2018, http://www.opiniomics.org/a-simple-test-for-uncorrected-insertions-and-deletions-indels-in-bacterial-genomes/). One way to rule out errors in gene models is to search against the six-frame translation of the genome, but six-frame searches are slow and cumbersome.

To make it easy to check for “missing” proteins, Curated BLAST searches against the six-frame translation of the genome. Because six-frame search takes several times longer than searching against the predicted proteins, Curated BLAST first searches against the predicted proteins and shows those results. It can then search against the six-frame translation while the user is inspecting the hits to the predicted proteins. Curated BLAST reports any hits to the six-frame translation that were not expected given the predicted proteins. For example, we recently found that the histidinol dehydrogenase of *Azospirillum brasilense* Sp245 was not annotated (see NCBI assembly ASM23736v1) because of a frameshift error in the genome sequence (10). As shown in Fig. 2, searching *A. brasilense* for “histidinol dehydrogenase” finds two nearby reading frames that are similar to characterized histidinol dehydrogenases. The two reading frames cover the N- and C-terminal portions of the protein (this can be seen by hovering on “62% id.” or “59% id.”).

We developed six-frame search with bacterial and archaeal genomes in mind, and it does not take splicing into account. Nevertheless, it may be useful for smaller eukaryotic genomes. Six-frame search is available for genomes of up to 30 Mb.

Integration with genome databases. Curated BLAST works with genomes from NCBI’s assembly database (11), from JGI’s Integrated Microbial Genomes (12), from MicrobesOnline (13), or from the Fitness Browser (14). Curated BLAST also works with UniProt proteomes (3), although six-frame search is not available for proteomes. For each of these sources, the user can search for the genome of interest by genus name, species name, and/or strain name. Alternatively, the user can upload protein or nucleotide sequences in FASTA format.

Sources of characterized proteins. Curated BLAST relies on eight databases for curated descriptions of the functions of characterized proteins:

- BRENDA, a database of enzymes (15).

TABLE 1 Sources of characterized proteins

Database	No. of entries	No. of distinct sequences
BRENDA	21,751	21,497
CAZy	8,878	8,629
CharProtDB	8,021	7,961
EcoCyc	4,161	4,108
Fitness Browser	1,321	1,319
MetaCyc	6,482	6,474
REBASE	3,447	2,749
Swiss-Prot	87,332	85,836
Total	141,393	115,874

- CAZy, a database of carbohydrate-active enzymes (16). Only the experimentally characterized proteins are included.
- CharProtDB, a database of characterized proteins (17).
- EcoCyc, a database of genes in *Escherichia coli* K-12 (18).
- MetaCyc, a database of metabolism and enzymes (19).
- REBASE, a database of DNA restriction and modification enzymes (20).
- Swiss-Prot, the manually curated section of UniProt (3). Only proteins with experimental evidence as to their function are included.
- The Fitness Browser, a database of genome-wide mutant fitness data and of proteins whose functions were identified from mutant phenotypes and comparative genomics (14).

Some of the proteins in CharProtDB, EcoCyC, and Swiss-Prot are not actually characterized and have vague annotations, but these vague annotations are unlikely to match a query, and so they rarely affect the results. Overall, Curated BLAST's database contains 115,874 different protein sequences, with most of these originating from Swiss-Prot (Table 1).

DISCUSSION

The most important limitation of Curated BLAST is the underlying database of characterized proteins. Although this database contains over 100,000 different characterized protein sequences, there are many more characterized proteins that have not been curated. Most of the microbial proteins that have been characterized in the last few years are probably missing (8). As natural language processing technology improves, it may become feasible to extract protein functions from the full text of papers along with sequence identifiers.

Another limitation of Curated BLAST is that a text query might not include all of the characterized proteins of interest. Except for enzymes, there is no shared nomenclature for protein functions across the curated databases. For enzymes, Curated BLAST works well with Enzyme Commission (EC) numbers as the queries if the "word search" option is selected. (Our database links over 45,000 different protein sequences to over 5,000 fully specific four-level EC numbers.) Even so, EC numbers might be overly specific if several similar reactions are of interest. For instance, oxidizing the same substrate with different electron acceptors corresponds to different EC numbers. Also, EC numbers can change, and some of the curated entries contain out-of-date EC numbers. The results page includes a link to the list of curated genes that match the query, which you can use to see if Curated BLAST is considering all of the activities that you expect.

Another limitation is that the activity of interest may be provided by a protein that is not related to any characterized protein that has the activity, even though the protein is similar to characterized proteins with similar activities. In this situation, searching for relevant protein families may be more effective than searching for the specific activity.

MATERIALS AND METHODS

Steps in curated BLAST. The steps in Curated BLAST are as follows:

- First, search for a genome of interest in one of the genome databases. Curated BLAST shows the user a list of genomes to choose from and a text box to enter a query. Given the selected genome, Curated BLAST fetches the predicted protein sequences and the genome sequence.
- Alternatively, the user can upload a genome or a proteome and enter a query.
- Either way, given the query, Curated BLAST finds curated proteins whose descriptions match. By default, a curated protein's description matches if it contains the query as a substring. Matching ignores the capitalization of the query or the description. Wild-card searches are also supported: for example, "chl%reduct" will match "Perchlorate reductase subunit alpha." If the user selects the "Match whole words only" option, then the query must be present as complete words (for instance, "chlorate" will match "Chlorate reductase" but not "Perchlorate reductase").
- Curated BLAST extracts the sequences of these curated proteins and uses Ublast (version 10.0) to compare these curated sequences to the predicted proteins, with a maximum E value of 0.01 (21).
- Curated BLAST sorts the hits by percent identity × percent coverage (highest first) and groups together hits for the same predicted protein. It shows a short description of each protein and the curated proteins to which it is similar (Fig. 1). If a protein has four or more hits, only the top three are shown initially, along with a link to view all of the hits for that protein.
- If the genome sequence is available and is under 30 Mb, Curated BLAST identifies all stretches of 30 amino acids or more without a stop codon and uses Ublast to compare these to the curated sequences. To avoid reporting hits against the six-frame translation that were already expected given the hits to the predicted proteins, hits for a given curated protein are ignored unless the score is noticeably better than the score of the hit to any predicted protein. The score is defined as percent identity of the alignment × percent coverage of the curated sequence. Noticeably better is defined as 1.1 × the best score to a predicted protein. Also, if the best hit for a reading frame is ignored, then all hits for that reading frame are ignored.
- Curated BLAST sorts the remaining hits for the six-frame translation by their best score and shows the hits for each reading frame.

Software. Curated BLAST for genomes is implemented in CGI (common gateway interface) scripts and Perl. It uses the same SQLite3 relational database as PaperBLAST (8).

Sources of characterized proteins. We previously described how EcoCyc and the characterized subset of Swiss-Prot were incorporated into the PaperBLAST database (8). For this study, we used EcoCyc 22.5 (downloaded in November 2018) and we downloaded Swiss-Prot on 29 October 2018. We also incorporated curated gene descriptions from BRENDA, CAZy, CharProtDB, MetaCyc, REBASE, and the Fitness Browser.

BRENDA was downloaded on 11 January 2018. Only entries that contain a UniProt identifier and at least one publication were retained. The corresponding protein sequences were obtained from UniProt. Protein sequences that were fragments (according to UniProt's FASTA files) were excluded.

CAZy was obtained in FASTA format and as a table of EC numbers (http://csbl.bmb.uga.edu/dbCAN/download/CAZyDB-ec-info.txt.07-20-2017) via dbCAN (22) on 15 November 2017. Only entries that are linked to EC numbers, which should be experimentally characterized, were retained. Entries whose description matched "frameshift" or "fragment" were excluded.

CharProtDB was downloaded on 10 April 2017. Entries of type trusted_uniprot were excluded as they are redundant with Swiss-Prot. Entries of type trusted_aspgd entries were excluded as they tended to have less specific functional information.

We used MetaCyc release 20.5 (downloaded on 24 April 2017) and parsed the protein table (proteins.dat). Only entries with UniProt identifiers and with a reference to at least one paper within the comment field were retained. The corresponding protein sequences were obtained from UniProt. Protein sequences that were fragments were excluded. Protein descriptions corresponding to the MetaCyc entry were obtained from the COMMON-NAME field or from the enzymatic reaction(s) that was linked to each entry. A few entries that lacked names and were not associated with enzymatic reactions were excluded.

REBASE was downloaded on 8 December 2017. REBASE entries were retained only if the sequence specificity of the methyltransferase or nuclease is known. Entries without links to papers were retained because these entries are also usually based on experimental evidence (and not just sequence similarity).

Curated reannotations of genes' functions were obtained from the Fitness Browser (http://fit.genomics.lbl.gov/) on 6 November 2018. (These reannotations are in the Reannotation table of the Fitness Browser's SQLite3 database, which is available for download [http://fit.genomics.lbl.gov/cgi_data/feba.db].) Each entry describes the putative function of a protein-encoding gene, as determined from its mutant phenotypes and comparative sequence analysis, along with a rationale. About half of the entries have been described in previous publications (10, 14).

PaperBLAST also includes another curated database, GeneRIF (23), which links papers to genes and includes a short summary of the findings about the gene(s). We did not include GeneRIF in Curated BLAST because most of the entries are not descriptions of a protein's function. For instance, many of the entries are about other aspects of genes such as expression patterns, and many of the entries mention more than one protein. If we had included entries for papers that GeneRIF links to just one protein, then the number of proteins with curated information would increase by 23% (from 115,874 to 149,694). The practical benefit would probably be modest, because 73% of these additional proteins are very similar (over 80% identity and over 75% coverage) to proteins that are curated by other resources.

Rationale for the scoring and ranking of hits. Curated BLAST uses a fixed E value threshold of 0.01. One limitation of Ublast is that it is not very sensitive to remote homologies (under 30% identity or so). This limitation is primarily due to Ublast's heuristics and not the E value cutoff. Because Ublast does not find remote homologs, we do not think it would be useful to change the E value threshold. Also, the distant homologies (that are missed by Ublast) are of questionable use for functional annotation anyway.

Although tools for homology search often rank the hits by the highest alignment score or bit score (or equivalently by the lowest E value), we thought it unwise to rank Curated BLAST's hits this way. The problem is that a moderate-percent-identity alignment to a long protein will have a higher bit score than a high-percent-identity similarity to a short protein. So, sorting by bit score would systematically bias the results to show homologs of longer proteins. (For tools that compare a genome to a single sequence, this issue would not arise.) Instead, we score by a combination of the percent identity and the percent coverage. A low percent identity often indicates that the proteins have different functions: for instance, enzymes with less than 40% identity are likely to act on different substrates (although the degree of conservation varies across families of enzymes) (24). And a low percent coverage implies that the proteins have differences in their domain content and are likely to have different functions (25).

Availability of data and source code. The data and the source code for the December 2018 release of Curated BLAST for Genomes and PaperBLAST are available at https://doi.org/10.6084/m9.figshare.7439216.v1. The latest version of the code is available at https://github.com/morgannprice/PaperBLAST.

ACKNOWLEDGMENT

This material by ENIGMA—Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory, is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research, under contract number DE-AC02-05CH11231.

REFERENCES

1. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5:e1000605. https://doi.org/10.1371/journal.pcbi.1000605.
2. Bender KS, Shang C, Chakraborty R, Belchik SM, Coates JD, Achenbach LA. 2005. Identification, characterization, and classification of genes encoding perchlorate reductase. J Bacteriol 187:5090–5096. https://doi.org/10.1128/JB.187.15.5090-5096.2005.
3. The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158–D169. https://doi.org/10.1093/nar/gkw1099.
4. Youngblut MD, Tsai C-L, Clark IC, Carlson HK, Maglaqui AP, Gau-Pan PS, Redford SA, Wong A, Tainer JA, Coates JD. 2016. Perchlorate reductase is distinguished by active site aromatic gate residues. J Biol Chem 291:9190–9202. https://doi.org/10.1074/jbc.M116.714618.
5. Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L. 2015. Update on RefSeq microbial genomes resources. Nucleic Acids Res 43:D599–D605. https://doi.org/10.1093/nar/gku1062.
6. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 42:D206–D214. https://doi.org/10.1093/nar/gkt1226.
7. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462. https://doi.org/10.1093/nar/gkv1070.
8. Price MN, Arkin AP. 2017. PaperBLAST: text mining papers for information about homologs. mSystems 2:e00039-17. https://doi.org/10.1128/mSystems.00039-17.
9. Venter E, Smith RD, Payne SH. 2011. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS One 6:e27587. https://doi.org/10.1371/journal.pone.0027587.
10. Price MN, Zane GM, Kuehl JV, Melnyk RA, Wall JD, Deutschbauer AM, Arkin AP. 2018. Filling gaps in bacterial amino acid biosynthesis pathways with high-throughput genetics. PLoS Genet 14:e1007147. https://doi.org/10.1371/journal.pgen.1007147.
11. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, Connor R, Fiorini N, Funk K, Hefferon T, Holmes JB, Kim S, Kimchi A, Kitts PA, Lathrop S, Lu Z, Madden TL, Marchler-Bauer A, Phan L, Schneider VA, Schoch CL, Pruitt KD, Ostell J. 2019. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 47:D23–D28. https://doi.org/10.1093/nar/gky1069.
12. Chen IA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R, Smirnova T, Kirton E, Jungbluth SP, Woyke T, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC. 2019. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res 47(D1): D666–D677. https://doi.org/10.1093/nar/gky901.
13. Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, Friedland GD, Huang KH, Keller K, Novichkov PS, Dubchak IL, Alm EJ, Arkin AP. 2010. MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res 38:D396–D400. https://doi.org/10.1093/nar/gkp919.
14. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, Kuehl JV, Melnyk RA, Lamson JS, Suh Y, Carlson HK, Esquivel Z, Sadeeshkumar H, Chakraborty R, Zane GM, Rubin BE, Wall JD, Visel A, Bristow J, Blow MJ, Arkin AP, Deutschbauer AM. 2018. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature 557:503–509. https://doi.org/10.1038/s41586-018-0124-0.
15. Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J, Schomburg D. 2017. BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Res 45:D380–D388. https://doi.org/10.1093/nar/gkw952.
16. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:D490–D495. https://doi.org/10.1093/nar/gkt1178.
17. Madupu R, Richter A, Dodson RJ, Brinkac L, Harkins D, Durkin S, Shrivastava S, Sutton G, Haft D. 2012. CharProtDB: a database of experimentally characterized protein annotations. Nucleic Acids Res 40:D237–D241. https://doi.org/10.1093/nar/gkr1133.
18. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD. 2005. EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res 33:D334–D337. https://doi.org/10.1093/nar/gki108.
19. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Paley S, Popescu L, Pujar A, Shearer AG, Zhang P, Karp PD. 2010. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 38: D473–D479. https://doi.org/10.1093/nar/gkp875.

20. Roberts RJ, Vincze T, Posfai J, Macelis D. 2015. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 43:D298–D299. https://doi.org/10.1093/nar/gku1046.

21. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. https://doi.org/10.1093/bioinformatics/btq461.

22. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40:W445–W451. https://doi.org/10.1093/nar/gks479.

23. Mitchell JA, Aronson AR, Mork JG, Folk LC, Humphrey SM, Ward JM. 2003. Gene indexing: characterization and analysis of NLM's GeneRIFs. AMIA Annu Symp Proc 2003:460–464.

24. Tian W, Skolnick J. 2003. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333:863–882. https://doi.org/10.1016/j.jmb.2003.08.057.

25. Galperin MY, Koonin EV. 1998. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In Silico Biol 1:55–67.