Physiological Reports Open Access

Physiological Reports ISSN 2051-817X

ORIGINAL RESEARCH

Sequence-based searching of custom proteome and transcriptome databases

Barbara Medvar[1,2,3], Abhijit Sarkar[1,2,3], Mark Knepper[1] & Trairak Pisitkun[1,4]

1 Epithelial Systems Biology Laboratory, Systems Biology Center, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland
2 Vitreous State Laboratory, The Catholic University of America, Washington, District of Columbia
3 Physics Department, The Catholic University of America, Washington, District of Columbia
4 Center of Excellence in Systems Biology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

Keywords
BLAST, database, IMCD, kidney.

Correspondence
Mark A. Knepper, National Institutes of Health, Bldg. 10, Room 6N307, 10 CENTER DR, MSC-1603, Bethesda, MD 20892-1603.
Tel: +1 (301) 496 3064
Fax: +1 (301) 402 1443
E-mail: knepperm@nhlbi.nih.gov

Funding Information
TP was supported by Chulalongkorn Academic Advancement into Its 2nd Century (CUAASC) Project and Thailand Research Fund (RSA5880014).

Received: 13 June 2018; Revised: 2 August 2018; Accepted: 6 August 2018

doi: 10.14814/phy2.13846

***Physiol Rep*, 6 (18), 2018, e13846,**
https://doi.org/10.14814/phy2.13846

Abstract

A long-term goal in renal physiology is to understand the mechanisms involved in collecting duct function and regulation at a cellular and molecular level. The first step in modeling of these mechanisms, which can provide a guide to experimentation, is the generation of a list of model components. We have curated a list of proteins expressed in the rat renal inner medullary collecting duct (IMCD) from proteomic data from 18 different publications. The database has been posted as a public resource at https://hpcwebapps.cit.nih.gov/ESBL/Database/IMCD_Proteome_Database/. It includes 8956 different proteins. To search the IMCD Proteomic Database efficiently, we have created a Java-based program called *curated database Basic Local Alignment Search Tool* (cdbBLAST), which uses the NCBI BLAST kernel to search for specific amino acid sequences corresponding to proteins in the database. cdbBLAST reports information on the matched protein and identifies proteins in the database that have similar sequences. We have also adapted cdbBLAST to interrogate our previously published IMCD Transcriptome Database. We have made the cdbBLAST program available for use either as a web application or a downloadable .jar file at https://hpcwebapps.cit.nih.gov/ESBL/Database/cdbBLAST/. Database searching based on protein sequence removes ambiguities arising from the standard search method based on official gene symbols and allows the user efficient identification of related proteins that may fulfill the same functional roles.

Introduction

With the advent of large-scale proteomic and transcriptomic experiments for profiling gene expression, data access and integration has become rate-limiting for acquisition of biological knowledge. Access is facilitated through the creation of databases of curated datasets. For example, our laboratory alone has generated approximately 70 such databases, accessible at https://hpcwebapps.cit.nih.gov/ESBL/Database/index.html. Because of the large amount of data within these types of databases, it becomes difficult to find information about a specific protein/transcript. The question then becomes: what is the best way to find a particular protein/transcript within the databases? A general strategy to database searches is to utilize so-called “keys” or “indices” that constitute finite search target lists. In systems biology, there are a few common “keys,” namely a common name for a protein/transcript, a protein/transcript’s gene symbol, or the amino acid/nucleotide sequence for a protein/transcript. Problems arise when a database is searched using a common name for the protein/transcript in question; common or large proteins have multiple names and can be difficult to find if the user is not searching for the name used within the database. Searching using a gene symbol is a better option, however it is not without its own difficulties. As with searching using a common name, a well-studied protein can be linked to multiple gene symbols. A second issue that arises from using the gene symbol as the "key" to searching databases is the inability to comprehensively search for related proteins. Unless all proteins share a similar gene symbol (e.g., Aqp1, Aqp2, etc.), it is extremely difficult to find all related proteins, for example, those with a particular domain. The best solution to the problems that arise from searching by gene symbol or a common name is to search curated databases using a protein/transcript's amino acid/nucleotide sequence. Each protein/transcript has a unique sequence, allowing the correct protein/transcript to be found regardless of the gene symbol or name used in the database. Using a sequence also allows for related proteins/transcripts to be found based on similar sequence matches.

The primary goal of this paper is to introduce a sequence-based search tool, called cdbBLAST (curated database BLAST), that uses NCBI's *blastp* kernel to uniquely find a protein/transcript within a curated database, as well as find similar proteins/transcripts, and provides information from the database on each of the proteins/transcripts in the results. Another benefit of cdbBLAST is that it allows for domain sequence searches, so it is possible to find all proteins/transcripts within a curated database that share a domain. We illustrate the use of this tool by applying it to two newly curated databases, a database of all rat inner medullary collecting duct transcripts, and a database of all rat inner medullary collecting duct proteins. The former is created by combining two published studies, one from RNA-seq and one from expression microarrays, while the latter is a created by distilling information from 18 published studies.

Methods

We curated a comprehensive list of proteins found in the rat IMCD cells from 18 studies (van Balkom et al. 2004; Hoffert et al. 2004, 2006, 2012, 2014; Barile et al. 2005; Hoorn et al. 2005; Pisitkun et al. 2006; Simons et al. 2006; Yu et al. 2006; Sachs et al. 2008; Bansal et al. 2010; Tchapyjnikov et al. 2010; Zhao et al. 2012; Bradford et al. 2014; Trepiccione et al. 2014; Pickering et al. 2016; LeMaire et al. 2017) done within our laboratory. Some of these proteomic studies include multiple experimental methods, which results in 20 "sources" describing which experimental method identified a specific protein. This database has been put together on a publicly accessible webpage (https://hpcwebapps.cit.nih.gov/ESBL/Database/IMCD_Proteome_Database/).

We have created a sequence-based search tool, called cdbBLAST, which allows databases to be searched for a certain protein using a protein-to-protein sequence comparison. cdbBLAST uses the NCBI *blastp* kernel to search the database for the protein in question and returns similar proteins that are found in the database as well. It also returns information from the database on each of the proteins in the results. cdbBLAST can be used as a web-based servlet or as a downloadable GUI.

When the user types (or pastes) a string of amino acids into the search box and hits submit, cdbBLAST turns that string into a .txt file, and uses it as input to the NCBI *blastp* kernel. The input needs to be a minimum of 11 amino acids in length. The blastp kernel then compares this input file to a database of sequences, and prints the results of the comparison (Fig. 1). The cdbBLAST web-based servlet then prints these results as a webpage. The cdbBLAST GUI version outputs the results directly in blastp format.

Software implementation

cdbBLAST, both the web-based servlets and the downloadable version, were written in Java (Java Development Kit 1.8 Update 121) using NetBeans IDE 8.1 as an integrated development environment. They were implemented using Apache Tomcat 8.0.15. The 2.4.0 version of *blastp* kernel from NCBI is used, and was downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+ as part of the blast 2.4.0+ folder.

cdbBLAST takes an amino acid sequence as input and compares it to a list of FASTA sequences corresponding to entries in the curated database. This list includes not only FASTA sequences and the typical FASTA metadata, but additional information from the databases that are built into the metadata.

These FASTA sequence lists were created using the Automated Bioinformatics Extractor (ABE) to convert the *RefSeq* protein accession numbers of the entries to FASTA amino acid sequence and metadata. The database information was then combined with the existing metadata for each protein/transcript. Next the list was copied into a Notepad file, and saved as a "filename.fasta" file. Subsequently, in the command line, the BLAST code makeblastdb.exe was run, using this new .fasta file as part of the input. This results in three new files being created: filename.fasta.phr, filename.fasta.pin, and filename.fasta.psq. These three files are required to run blastp, and thus cdbBLAST (Fig. 2). A more thorough step by step process can be found in the Appendix, downloadable from https://hpcwebapps.cit.nih.gov/ESBL/Database/cdbBLAST-Appendix/Appendix.html.

Results

We have curated a rat IMCD proteome database from 18 studies produced in our laboratory, resulting in a database of 8956 proteins. This database has been made into a publicly accessible webpage (https://hpcwebapps.cit.nih.gov/ESBL/Database/IMCD_Proteome_Database/) (Fig. 3).

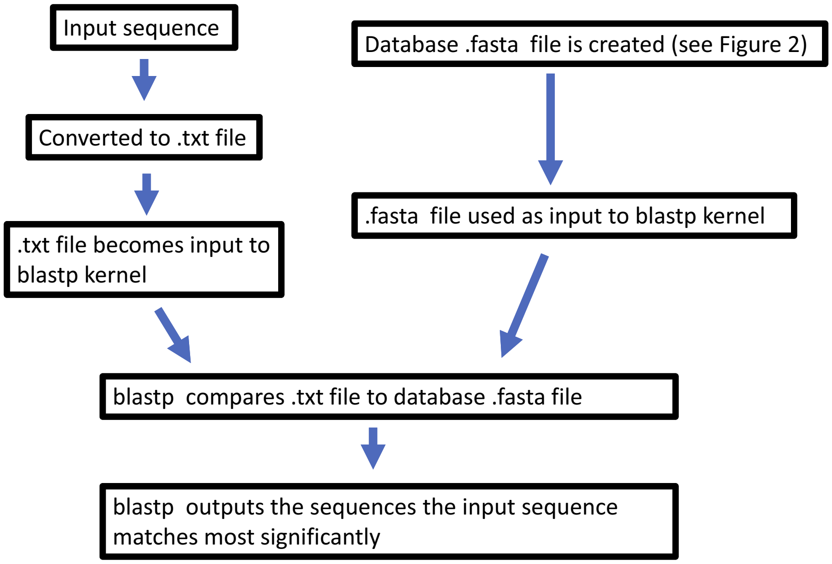

Input sequence
Converted to .txt file
.txt file becomes input to blastp kernel
Database .fasta file is created (see Figure 2)
.fasta file used as input to blastp kernel
blastp compares .txt file to database .fasta file
blastp outputs the sequences the input sequence matches most significantly

Figure 1. A flowchart explaining the algorithm of cdbBLAST. The input sequence is converted to a .txt file, which becomes the input to the BLAST kernel. This input file is compared to a .fasta database, and the sequences in the .fasta database that match most closely to the input sequence are output in the results.

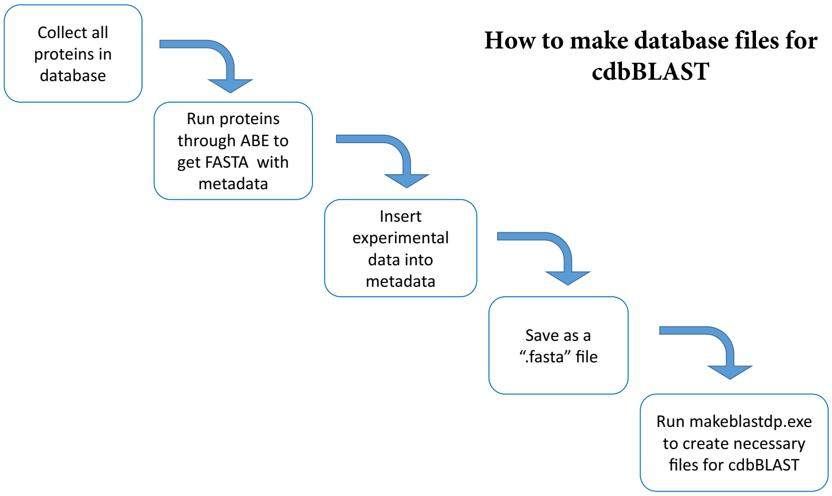
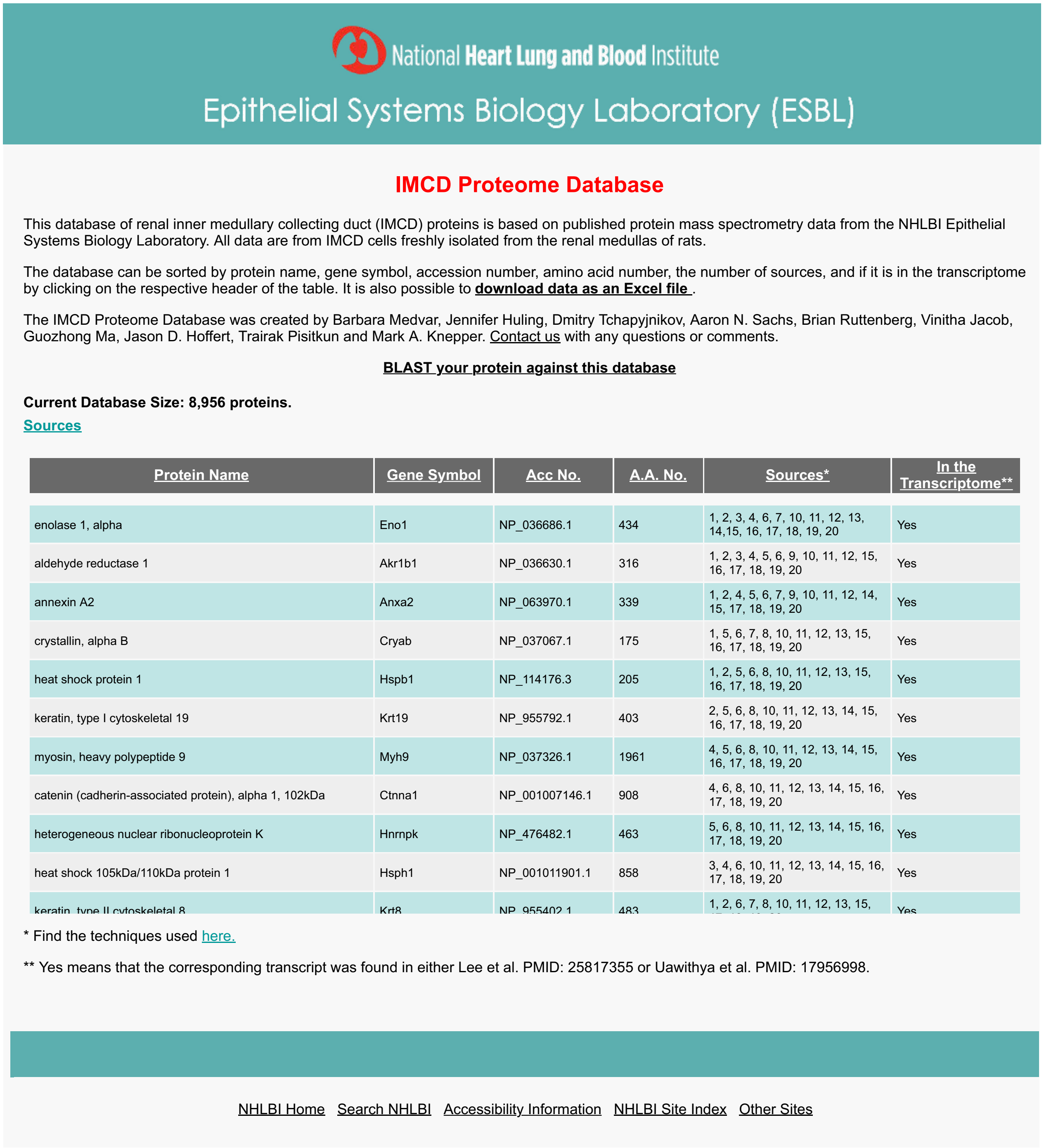

How to make database files for cdbBLAST
Collect all proteins in database
Run proteins through ABE to get FASTA with metadata
Insert experimental data into metadata
Save as a ".fasta" file
Run makeblastdp.exe to create necessary files for cdbBLAST

Figure 2. A flowchart explaining the process for creating a .fasta database to be used with cdbBLAST. We start with a list of all proteins* in the database. That list is then run through Automated Bioinformatics Extractor (ABE) to get the FASTA sequence and metadata associated with each protein in the list. Next, experimental information is added to the metadata for each protein, and this new list is saved as a ".fasta" file. The .fasta file is then the input for makeblastdb.exe, which creates the three necessary files needed to run cdbBLAST. A more detailed explanation can be found in the Appendix, downloadable from https://hpcwebapps.cit.nih.gov/ESBL/Database/cdbBLAST-Appendix/Appendix.html. *This figure describes proteins being made into a database. The process for creating a database of transcripts is the same.

This database shows how many different experimental datasets each protein was found in, as well as its protein name, official gene symbol, accession number, and number of amino acids. The database also shows, and can be sorted by, whether the corresponding mRNA is present in the transcriptome.

To best search large databases like this rat IMCD proteome, we have created a sequence-based search tool called cdbBLAST. cdbBLAST allows the user to input a full amino acid sequence, or a partial sequence, and compare it against a list of FASTA sequences corresponding to the proteins/transcripts in the curated database. We have created a rat IMCD proteome database and combined two existing rat IMCD transcriptome databases to be searched using cdbBLAST, which has been made publicly accessible (https://hpcwebapps.cit.nih.gov/ESBL/Database/cdbBLAST) as well as downloadable as a .jar file (Fig. 4).



National Heart Lung and Blood Institute

Epithelial Systems Biology Laboratory (ESBL)

IMCD Proteome Database

This database of renal inner medullary collecting duct (IMCD) proteins is based on published protein mass spectrometry data from the NHLBI Epithelial Systems Biology Laboratory. All data are from IMCD cells freshly isolated from the renal medullas of rats.

The database can be sorted by protein name, gene symbol, accession number, amino acid number, the number of sources, and if it is in the transcriptome by clicking on the respective header of the table. It is also possible to **download data as an Excel file** .

The IMCD Proteome Database was created by Barbara Medvar, Jennifer Huling, Dmitry Tchapyjnikov, Aaron N. Sachs, Brian Ruttenberg, Vinitha Jacob, Guozhong Ma, Jason D. Hoffert, Trairak Pisitkun and Mark A. Knepper. Contact us with any questions or comments.

BLAST your protein against this database

Current Database Size: 8,956 proteins.

Sources

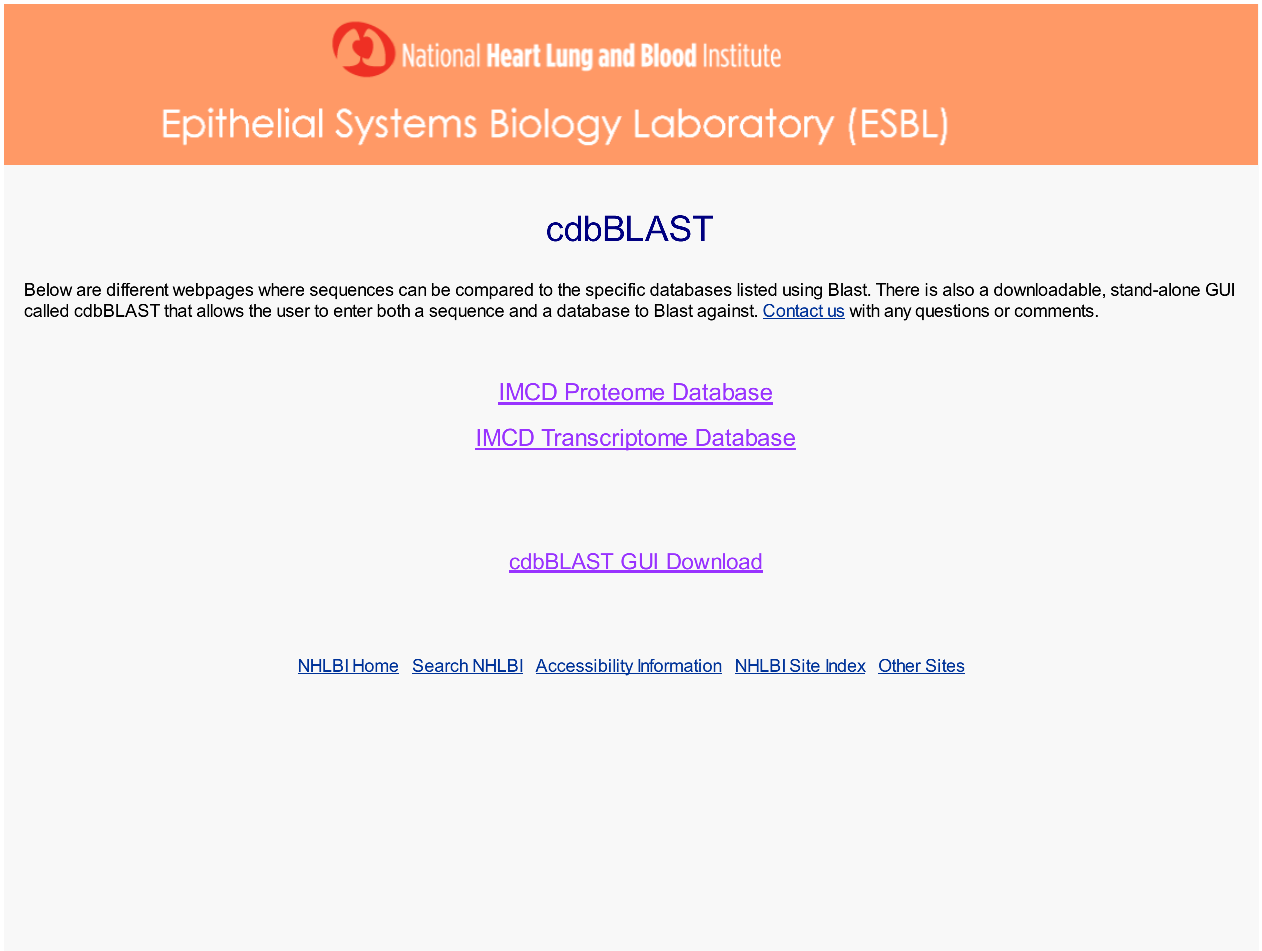
Protein Name	Gene Symbol	Acc No.	A.A. No.	Sources*	In the Transcriptome**
enolase 1, alpha	Eno1	NP_036686.1	434	1, 2, 3, 4, 6, 7, 10, 11, 12, 13, 14,15, 16, 17, 18, 19, 20	Yes
aldehyde reductase 1	Akr1b1	NP_036630.1	316	1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 15, 16, 17, 18, 19, 20	Yes
annexin A2	Anxa2	NP_063970.1	339	1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 14, 15, 17, 18, 19, 20	Yes
crystallin, alpha B	Cryab	NP_037067.1	175	1, 5, 6, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20	Yes
heat shock protein 1	Hspb1	NP_114176.3	205	1, 2, 5, 6, 8, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20	Yes
keratin, type I cytoskeletal 19	Krt19	NP_955792.1	403	2, 5, 6, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	Yes
myosin, heavy polypeptide 9	Myh9	NP_037326.1	1961	4, 5, 6, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	Yes
catenin (cadherin-associated protein), alpha 1, 102kDa	Ctnna1	NP_001007146.1	908	4, 6, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	Yes
heterogeneous nuclear ribonucleoprotein K	Hnrnpk	NP_476482.1	463	5, 6, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	Yes
heat shock 105kDa/110kDa protein 1	Hsph1	NP_001011901.1	858	3, 4, 6, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	Yes
keratin, type II cytoskeletal 8	Krt8	NP_955402.1	483	1, 2, 6, 7, 8, 10, 11, 12, 13, 15,	Yes

* Find the techniques used here.

** Yes means that the corresponding transcript was found in either Lee et al. PMID: 25817355 or Uawithya et al. PMID: 17956998.

Figure 3. A publicly available webpage was created to provide users with a complete list of proteins found in the rat IMCD, as identified in 18 studies. Access to this webpage can be found at https://hpcwebapps.cit.nih.gov/ESBL/Database/IMCD_Proteome_Database/. There is a link to a list of all sources, as well as a link to download the data to an Excel file. The data on this webpage can be sorted by the number of sources a protein is found in, amino acid number, if a protein is found in the transcriptome, or alphabetically by gene symbol, protein name, or RefSeq number.

The proteome cdbBLAST search (Fig. 5A) allows the user to input an amino acid sequence and search the curated rat IMCD proteome database introduced in this paper. There is a link on both the main page and the results page of cdbBLAST to the proteome database. At the bottom of the search page, there is a set of criteria (e.g., expect threshold, word size, substitution matrix, and gap costs) the user can use to manipulate the results, based on NCBI's *BLAST* code (Altschul et al. 1990) (Fig. 5B). Once the user is content with the criteria and the sequence used, hitting "submit" will allow the code to run and brings up a results page. This results page shows a table at the top of the page with the proteins that match best to the search sequence. It also shows the related proteins' RefSeq number, which is a link to that protein in PubMed (National Center for Biotechnology Information, 2017), and the "score" and "e-value" as determined by NCBI's *blastp* code. The "score" is a normalized score for aligning pairs of residues between the query and the database sequences, where the "e-value" or expected value is the number of hits one can expect to see by chance. The higher the score, and the lower the e-value, the more significant the match. Below this table, each related protein is compared to the search sequence. Each comparison is labeled with the protein's RefSeq number, protein name, and what sources from the IMCD proteome that protein was found in. Between the table and the comparisons is a link labeled "Sources listed here" which allows the user to open the "Sources" page from the IMCD proteome database in another tab so they are able to see what each numbered source is without losing their cdbBLAST search results. To show an example of what the results look like, we used the partial sequence associated with the HECTc domain from Nedd4l as a sample query. Nedd4l is an E3 ubiquitin ligase most likely to ubiquitinate the water channel aquaporin-2 (Medvar et al. 2016) in IMCD cells. One use of cdbBLAST is searching a large-scale database



National Heart Lung and Blood Institute

Epithelial Systems Biology Laboratory (ESBL)

cdbBLAST

Below are different webpages where sequences can be compared to the specific databases listed using Blast. There is also a downloadable, stand-alone GUI called cdbBLAST that allows the user to enter both a sequence and a database to Blast against. Contact us with any questions or comments.

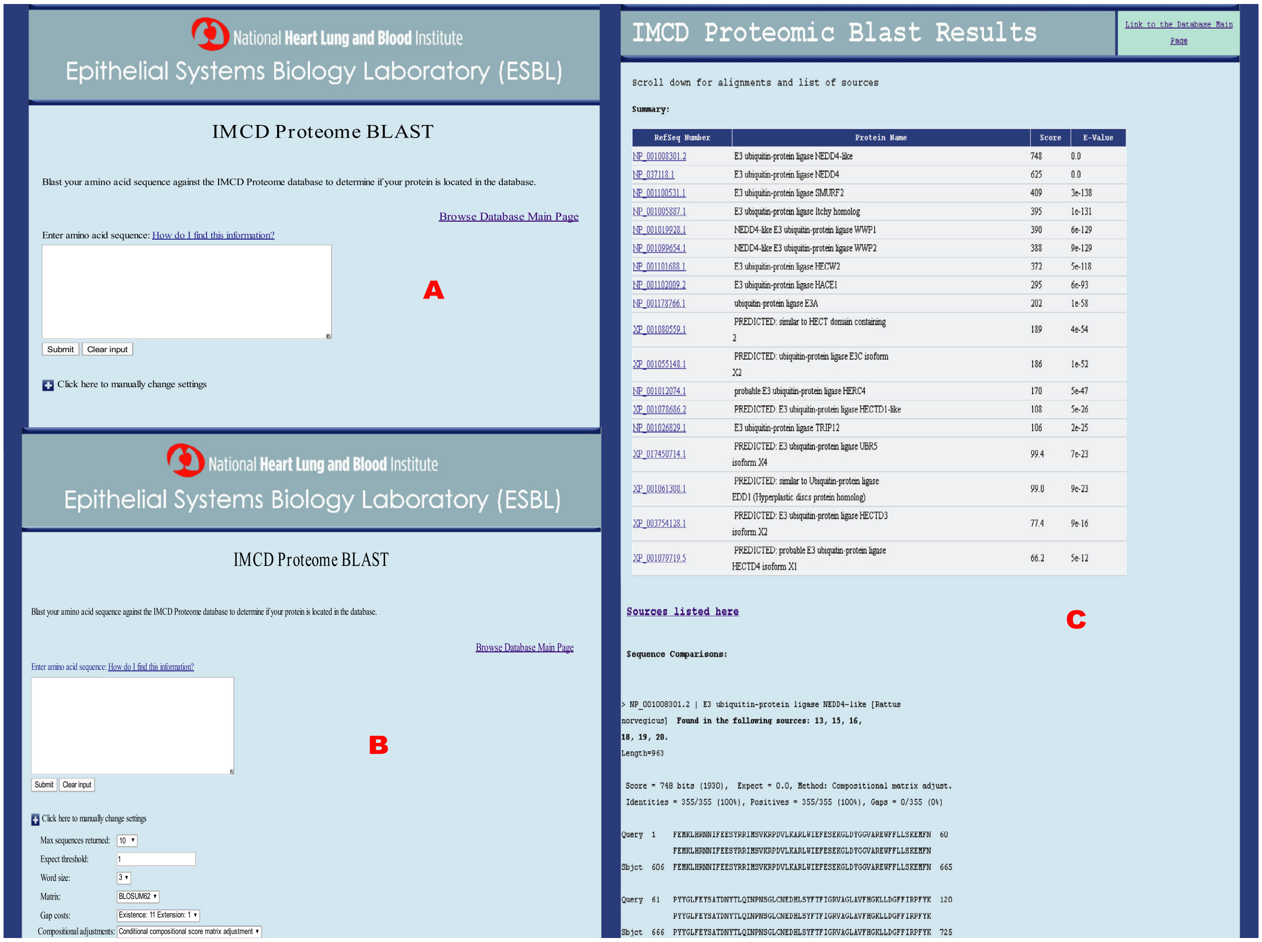
IMCD Proteome Database

IMCD Transcriptome Database

cdbBLAST GUI Download

NHLBI Home Search NHLBI Accessibility Information NHLBI Site Index Other Sites

Figure 4. A screenshot of the main cdbBLAST webpage. This page has links to the IMCD Proteome and Transcriptome cdbBLAST searches, which compare an input sequence to our databases of proteins and transcripts, respectively. There is also a link to the downloadable GUI version of cdbBLAST, where the user can compare a sequence of interest to their own database of sequences.


National Heart Lung and Blood Institute
Epithelial Systems Biology Laboratory (ESBL)
Submit
Clear input
A
B
C

IMCD Proteome BLAST

Blast your amino acid sequence against the IMCD Proteome database to determine if your protein is located in the database.

Browse Database Main Page

Enter amino acid sequence: How do I find this information?

Click here to manually change settings

IMCD Proteome BLAST

Blast your amino acid sequence against the IMCD Proteome database to determine if your protein is located in the database.

Browse Database Main Page

Enter amino acid sequence: How do I find this information?

Click here to manually change settings

Max sequences returned: 10

Expect threshold: 1

Word size: 3

Matrix: BLOSUM62

Gap costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

IMCD Proteomic Blast Results

Link to the Database Main Page

Scroll down for alignments and list of sources

Summary:

RefSeq Number	Protein Name	Score	E-Value
NP_001008301.2	E3 ubiquitin-protein ligase NEDD4-like	748	0.0
NP_037118.1	E3 ubiquitin-protein ligase NEDD4	625	0.0
NP_001100531.1	E3 ubiquitin-protein ligase SMURF2	409	3e-138
NP_001005887.1	E3 ubiquitin-protein ligase Itchy homolog	395	1e-131
NP_001019928.1	NEDD4-like E3 ubiquitin-protein ligase WWP1	390	6e-129
NP_001099654.1	NEDD4-like E3 ubiquitin-protein ligase WWP2	388	9e-129
NP_001101688.1	E3 ubiquitin-protein ligase HECW2	372	5e-118
NP_001102009.2	E3 ubiquitin-protein ligase HACE1	295	6e-93
NP_001178766.1	ubiquitin-protein ligase E3A	202	1e-58
XP_001080559.1	PREDICTED: similar to HECT domain containing 2	189	4e-54
XP_001055148.1	PREDICTED: ubiquitin-protein ligase E3C isoform X2	186	1e-52
NP_001012074.1	probable E3 ubiquitin-protein ligase HERC4	170	5e-47
XP_001078686.2	PREDICTED: E3 ubiquitin-protein ligase HECTD1-like	108	5e-26
NP_001026829.1	E3 ubiquitin-protein ligase TRIP12	106	2e-25
XP_017450714.1	PREDICTED: E3 ubiquitin-protein ligase UBR5 isoform X4	99.4	7e-23
XP_001061308.1	PREDICTED: similar to Ubiquitin-protein ligase EDD1 (Hyperplastic discs protein homolog)	99.0	9e-23
XP_003754128.1	PREDICTED: E3 ubiquitin-protein ligase HECTD3 isoform X2	77.4	9e-16
XP_001079719.5	PREDICTED: probable E3 ubiquitin-protein ligase HECTD4 isoform X1	66.2	5e-12

Sources listed here

Sequence Comparisons:

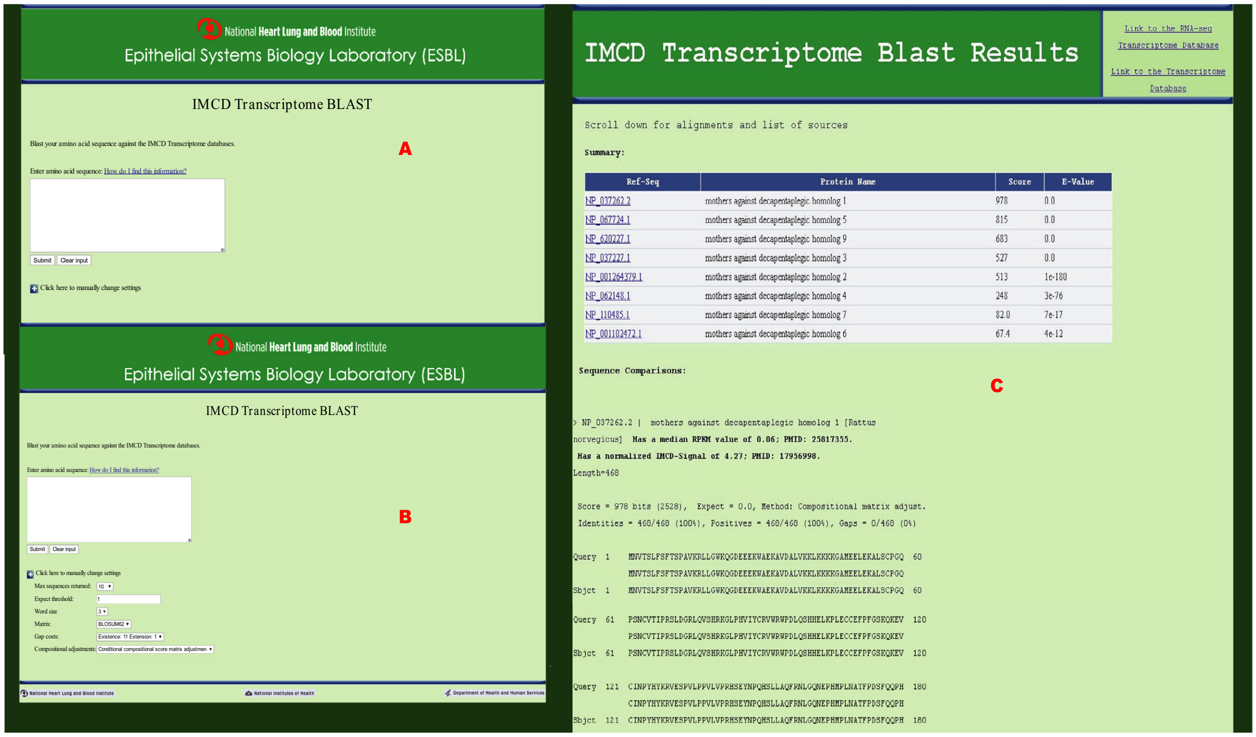
```
> NP_001008301.2 | E3 ubiquitin-protein ligase NEDD4-like [Rattus
norvegicus]  Found in the following sources: 13, 15, 16,
18, 19, 20.
Length=963

 Score = 748 bits (1930),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 355/355 (100%), Positives = 355/355 (100%), Gaps = 0/355 (0%)

Query  1    FEMKLHRNNIFEESYRRIMSVKRPDVLKARLWIEFESEKGLDYGGVAREWFFLLSKEMFN  60
            FEMKLHRNNIFEESYRRIMSVKRPDVLKARLWIEFESEKGLDYGGVAREWFFLLSKEMFN
Sbjct  606  FEMKLHRNNIFEESYRRIMSVKRPDVLKARLWIEFESEKGLDYGGVAREWFFLLSKEMFN  665

Query  61   PYYGLFEYSATDNYTLQINPNSGLCNEDHLSYFTFIGRVAGLAVFHGKLLDGFFIRPFYK  120
            PYYGLFEYSATDNYTLQINPNSGLCNEDHLSYFTFIGRVAGLAVFHGKLLDGFFIRPFYK
Sbjct  666  PYYGLFEYSATDNYTLQINPNSGLCNEDHLSYFTFIGRVAGLAVFHGKLLDGFFIRPFYK  725
```

Figure 5. A screenshot of the rat IMCD Proteomic cdbBLAST (A) initial page where a sequence can be typed or pasted into the white box, (B) settings that can be changed to alter the results of a search, and (C) results, using Nedd4l as a sample amino acid input sequence. The results include a summary table of the proteins whose sequences best match the query (input) sequence, which includes links to their PubMed protein pages. Below this table are comparisons of these top proteins to the query sequence. Within each comparison, the sources in which that top protein is found are listed. There is a link to the original database on both the initial page and the results page.

for an amino acid sequence associated with a specific domain to find all proteins within that domain. cdbBLAST is a useful tool for this type of search because of its ability to give not only results that match exactly, but also results that are similar. As seen in Figure 5C, using the HECTc domain of Nedd4l shows results not only for Nedd4l, but also for 17 similar proteins. For Nedd4l you can also see the beginning of the sequence comparison, as well as the datasets Nedd4l was found in.

The rat IMCD transcriptome cdbBLAST search (Fig. 6A) is very similar to the proteome search. However, this database is a combination of two different transcriptomic databases: one using RNA-sequencing (Lee et al. 2015) and the other using Affymetrix techniques (Uawithya et al. 2008). The results from this cdbBLAST search are structured the same as the proteome cdbBLAST search (see above), however the comparisons are labeled with the protein's RefSeq number, protein name, and the median RPKM value and/or the normalized IMCD-Signal (depending what study that protein is found in). As an example, we searched the IMCD transcriptome using the amino acid sequence for Smad1. Smad1 is a transcription factor found in the IMCD transcriptome as seen in Figure 6C. cdbBLAST also found many very similar proteins as well, some so similar they have the same e-value as Smad1. The comparison of the query to the database shows that Smad1 was found in both transcriptomic studies used to create this database.

The downloadable version of cdbBLAST (Fig. 7) is a slightly simpler version than the web servlets. It does not allow the user to adjust the settings and there are no clickable links in the results, but the information provided is the same as the cdbBLAST web-based searches, with there being a table of similar proteins followed by


A
B
C
National Heart Lung and Blood Institute
Submit
Clear input
National Heart Lung and Blood Institute
National Institutes of Health
Department of Health and Human Services

Epithelial Systems Biology Laboratory (ESBL)

IMCD Transcriptome BLAST

Blast your amino acid sequence against the IMCD Transcriptome databases.

Enter amino acid sequence: How do I find this information?

Click here to manually change settings

Epithelial Systems Biology Laboratory (ESBL)

IMCD Transcriptome BLAST

Blast your amino acid sequence against the IMCD Transcriptome databases.

Enter amino acid sequence: How do I find this information?

Click here to manually change settings

Max sequences returned:	10 ▾
Expect threshold:	1
Word size	3 ▾
Matrix:	BLOSUM62 ▾
Gap costs:	Existence: 11 Extension: 1 ▾
Compositional adjustments:	Conditional compositional score matrix adjustmen ▾

IMCD Transcriptome Blast Results

Link to the RNA-seq Transcriptome Database

Link to the Transcriptome Database

Scroll down for alignments and list of sources

Summary:

Ref-Seq	Protein Name	Score	E-Value
NP_037262.2	mothers against decapentaplegic homolog 1	978	0.0
NP_067724.1	mothers against decapentaplegic homolog 5	815	0.0
NP_620227.1	mothers against decapentaplegic homolog 9	683	0.0
NP_037227.1	mothers against decapentaplegic homolog 3	527	0.0
NP_001264379.1	mothers against decapentaplegic homolog 2	513	1e-180
NP_062148.1	mothers against decapentaplegic homolog 4	248	3e-76
NP_110485.1	mothers against decapentaplegic homolog 7	82.0	7e-17
NP_001102472.1	mothers against decapentaplegic homolog 6	67.4	4e-12

Sequence Comparisons:

```
> NP_037262.2 |  mothers against decapentaplegic homolog 1 [Rattus
norvegicus]  Has a median RPKM value of 0.06; PMID: 25817355.
 Has a normalized IMCD-Signal of 4.27; PMID: 17956998.
Length=468

 Score = 978 bits (2528),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 468/468 (100%), Positives = 468/468 (100%), Gaps = 0/468 (0%)

Query  1    MNVTSLFSFTSPAVKRLLGWKQGDEEEKWAEKAVDALVKKLKKKKGAMEELEKALSCPGQ  60
            MNVTSLFSFTSPAVKRLLGWKQGDEEEKWAEKAVDALVKKLKKKKGAMEELEKALSCPGQ
Sbjct  1    MNVTSLFSFTSPAVKRLLGWKQGDEEEKWAEKAVDALVKKLKKKKGAMEELEKALSCPGQ  60

Query  61   PSNCVTIPRSLDGRLQVSHRKGLPHVIYCRVWRWPDLQSHHELKPLECCEFPFGSKQKEV  120
            PSNCVTIPRSLDGRLQVSHRKGLPHVIYCRVWRWPDLQSHHELKPLECCEFPFGSKQKEV
Sbjct  61   PSNCVTIPRSLDGRLQVSHRKGLPHVIYCRVWRWPDLQSHHELKPLECCEFPFGSKQKEV  120

Query  121  CINPYHYKRVESPVLPPVLVPRHSEYNPQHSLLAQFRNLGQNEPHMPLNATFPDSFQQPH  180
            CINPYHYKRVESPVLPPVLVPRHSEYNPQHSLLAQFRNLGQNEPHMPLNATFPDSFQQPH
Sbjct  121  CINPYHYKRVESPVLPPVLVPRHSEYNPQHSLLAQFRNLGQNEPHMPLNATFPDSFQQPH  180
```

Figure 6. A screenshot of the rat IMCD Transcriptome cdbBLAST (A) initial page where a sequence can be typed or pasted into the white box, (B) settings that can be changed to alter the results of a search, and (C) results, using Smad1 as a sample amino acid input sequence. There are links to the original databases, a summary table of the transcripts whose sequences best match the query (input) sequence which includes links to their PubMed protein pages, and comparisons of these top transcripts to the query sequence. Each comparison also reports the median RPKM value and/or the normalized IMCD-Signal (depending what study that protein is found in) for that transcript.

sequence to sequence comparisons with database data. However, unlike the web servlets, the downloadable version of cdbBLAST can be used with any database of proteins or transcripts that has been correctly formatted. This allows users to integrate any prior data they wish into the database (and thus the output from the search), as well as allows the user to submit multiple sequences as input in a text file. A manual has been created to provide the user with a step-by-step guide to this process (see Appendix, download from https://hpcwebapps.cit.nih.gov/ESBL/Database/cdbBLAST-Appendix/Appendix.html). While this is the best use of the cdbBLAST GUI, we have also made the databases created in our laboratory available for download as well on the GUI download page.

Discussion

Our laboratory has created rat IMCD cdbBLAST searches for the proteome and transcriptome both as web-based servlets. We have also created a downloadable GUI version, with downloadable versions of our IMCD proteome and transcriptome databases available, as well as a manual explaining how the user can create a database from their own data.

cdbBLAST is a valuable tool that allows scientists to use the most unique search parameter when looking through a database: the amino acid/nucleotide sequence. With large databases created from large-scale proteomic and transcriptomic experiments becoming more common, a lot of useful data is being reported but not utilized due to the sheer volume of information available. As more is learned about the nature and function of specific proteins or transcripts, they are given new names or gene symbols. This makes older databases tedious and time consuming to search if they cannot be searched via amino acid/nucleotide sequences. There is no longer a need to do multiple searches within a database using every gene symbol from discovery to present in hopes of finding data on a specific protein/transcript.

cdbBLAST also allows the user to search for pieces of sequence, such as protein domains, allowing researchers to learn more about a group of proteins/transcripts in a single search. However, cdbBLAST requires at least 11 amino acids/nucleotides for the search to run, making it difficult search for smaller domains or motifs.

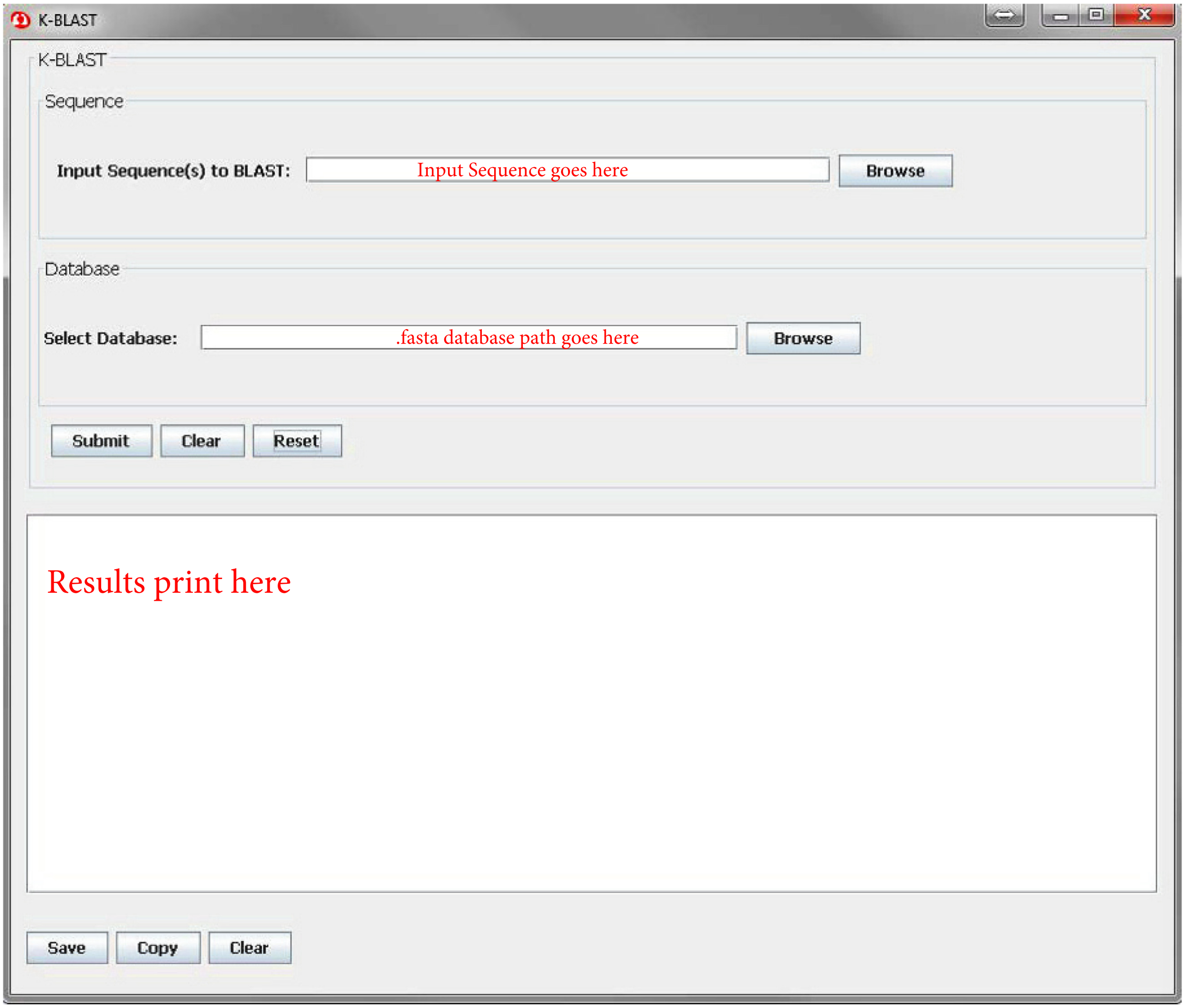


Figure 7. A screenshot of the downloadable version of cdbBLAST. Users can either type/paste in a sequence into the "Input Sequence to BLAST:" box or browse their computer for a .txt file to input. Users then browse their computer (or type in the path) for the .fasta database they either created (Fig. 2) or downloaded from our webpage before hitting "Submit". "Clear" clears only the input, while "Reset" clears the database and output sections as well. The results print similarly to the web-servlets, without any links to webpages. The results can be saved, copied, or cleared by hitting the corresponding buttons under the results box.

For every searched sequence using cdbBLAST, the results include proteins/transcripts that are similar to the search sequence. This makes it easier for researchers to find a related protein/transcript if the search protein/transcript is not found, or when the searched sequence is a different species than that of the database. Every protein/transcript shown in the results is accompanied by the database information associated with it, as well as a sequence comparison to the searched sequence, so the user can see how well the sequences match. For example, we have combined 18 published studies from our laboratory into one large IMCD proteome database of 8956 proteins. This database lists each proteins name, gene symbol, amino acid number, accession number, whether it is found in the transcriptome, and which of the sources (experiments from within these studies) that protein was found in. When this database is searched using cdbBLAST, each protein in the results lists the sources in which that protein was found. It is logical that the more times a protein was identified, the more likely it is to exist in the IMCD. This added information to the results of the search allows the user to see the relevant information without having to go back to the original database.

The web-based servlets we have created using cdbBLAST focus on the rat IMCD transcriptome and proteome. Specifically, they focus on the level of expression of the transcripts in the IMCD and what proteins are found in IMCD cells, respectively. The downloadable version of cdbBLAST allows the user to create their own searchable database(s) which can include any and all information they wish, from multiple datasets. Our transcriptome and proteome cdbBLAST databases are also available to be downloaded and used with this version. While amino acid/nucleotide sequences remain consistent through time, more information is learned about proteins and transcripts constantly. With the downloadable version of cdbBLAST, new information can be added to an existing searchable database with ease (see Appendix, download from https://hpcwebapps.cit.nih.gov/ESBL/Database/cdbBLAST-Appendix/Appendix.html).

Conflict of Interest

None declared.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

van Balkom, B. W., J. D. Hoffert, C. L. Chou, and M. A. Knepper. 2004. Proteomic analysis of long-term vasopressin action in the inner medullary collecting duct of the Brattleboro rat. Am. J. Physiol. Renal Physiol. 286:F216–F224.

Bansal, A. D., J. D. Hoffert, T. Pisitkun, S. Hwang, C. L. Chou, E. S. Boja, et al. 2010. Phosphoproteomic profiling reveals vasopressin-regulated phosphorylation sites in collecting duct. J. Am. Soc. Nephrol. 21:303–315.

Barile, M., T. Pisitkun, M. J. Yu, C. L. Chou, M. J. Verbalis, R. F. Shen, et al. 2005. Large scale protein identification in intracellular aquaporin-2 vesicles from renal inner medullary collecting duct. Mol. Cell Proteomics 4:1095–1106.

Bradford, D., V. Raghuram, J. L. Wilson, C. L. Chou, J. D. Hoffert, M. A. Knepper, et al. 2014. Use of LC-MS/MS and Bayes' theorem to identify protein kinases that phosphorylate aquaporin-2 at Ser256. Am. J. Physiol. Cell Physiol. 307:C123–C139.

Hoffert, J. D., B. W. van Balkom, C. L. Chou, and M. A. Knepper. 2004. Application of difference gel electrophoresis to the identification of inner medullary collecting duct proteins. Am. J. Physiol. Renal Physiol. 286:F170–F179.

Hoffert, J. D., T. Pisitkun, G. Wang, R. F. Shen, and M. A. Knepper. 2006. Quantitative phosphoproteomics of vasopressin-sensitive renal cells: regulation of aquaporin-2 phosphorylation at two sites. Proc. Natl Acad. Sci. USA 103:7159–7164.

Hoffert, J. D., T. Pisitkun, F. Saeed, J. H. Song, C. L. Chou, and M. A. Knepper. 2012. Dynamics of the G protein-coupled vasopressin V2 receptor signaling network revealed by quantitative phosphoproteomics. Mol. Cell Proteomics 11:M111 014613.

Hoffert, J. D., T. Pisitkun, F. Saeed, J. L. Wilson, and M. A. Knepper. 2014. Global analysis of the effects of the V2 receptor antagonist satavaptan on protein phosphorylation in collecting duct. Am. J. Physiol. Renal Physiol. 306:410–421.

Hoorn, E. J., J. D. Hoffert, and M. A. Knepper. 2005. Combined proteomics and pathways analysis of collecting duct reveals a protein regulatory network activated in vasopressin escape. J. Am. Soc. Nephrol. 16:2852–2863.

Lee, J. W., C. L. Chou, and M. A. Knepper. 2015. Deep sequencing in microdissected renal tubules identifies nephron segment-specific transcriptomes. J. Am. Soc. Nephrol. 26:2669–2677.

LeMaire, S. M., V. Raghuram, C. R. Grady, C. M. Pickering, C. L. Chou, E. N. Umejiego, et al. 2017. Serine/threonine phosphatases and aquaporin-2 regulation in renal collecting duct. Am. J. Physiol. Renal Physiol. 312:F84–F95.

Medvar, B., V. Raghuram, T. Pisitkun, A. Sarkar, and M. A. Knepper. 2016. Comprehensive database of human E3 ubiquitin ligases: application to aquaporin-2 regulation. Physiol. Genomics 48:502–512.

National Center for Biotechnology Information (NCBI) [Internet]. 2017. Bethesda (MD): National Library of Medicine (US) NCfBI. https://www.ncbi.nlm.nih.gov/. [October 1, 2017].

Pickering, C. M., C. Grady, B. Medvar, M. Emamian, P. C. Sandoval, Y. Zhao, et al. 2016. Proteomic profiling of nuclear fractions from native renal inner medullary collecting duct cells. Physiol. Genomics 48:154–166.

Pisitkun, T., J. Bieniek, D. Tchapyjnikov, G. Wang, W. W. Wu, R. F. Shen, et al. 2006. High-throughput identification of IMCD proteins using LC-MS/MS. Physiol. Genomics 25:263–276.

Sachs, A. N., T. Pisitkun, J. D. Hoffert, M. J. Yu, and M. A. Knepper. 2008. LC-MS/MS analysis of differential centrifugation fractions from native inner medullary collecting duct of rat. Am. J. Physiol. Renal Physiol. 295: F1799–F1806.

Simons, B. L., G. Wang, R. F. Shen, and M. A. Knepper. 2006. In vacuo isotope coded alkylation technique (IVICAT); an N-terminal stable isotopic label for quantitative liquid chromatography/mass spectrometry proteomics. Rapid Commun. Mass Spectrom. 20:2463–2477.

Tchapyjnikov, D., Y. Li, T. Pisitkun, J. D. Hoffert, M. J. Yu, and M. A. Knepper. 2010. Proteomic profiling of nuclei from native renal inner medullary collecting duct cells using LC-MS/MS. Physiol. Genomics 40:167–183.

Trepiccione, F., T. Pisitkun, J. D. Hoffert, S. B. Poulsen, G. Capasso, S. Nielsen, et al. 2014. Early targets of lithium in rat kidney inner medullary collecting duct include p38 and ERK1/2. Kidney Int. 86:757–767.

Uawithya, P., T. Pisitkun, B. E. Ruttenberg, and M. A. Knepper. 2008. Transcriptional profiling of native inner medullary collecting duct cells from rat kidney. Physiol. Genomics 32:229–253.

Yu, M. J., T. Pisitkun, G. Wang, R. F. Shen, and M. A. Knepper. 2006. LC-MS/MS analysis of apical and basolateral plasma membranes of rat renal collecting duct cells. Mol. Cell Proteomics 5:2131–2145.

Zhao, B., M. A. Knepper, C. L. Chou, and T. Pisitkun. 2012. Large-scale phosphotyrosine proteomic profiling of rat renal collecting duct epithelium reveals predominance of proteins involved in cell polarity determination. Am. J. Physiol. Cell Physiol. 302:C27–C45.