**Homology Service Product Description**

**Summary:**

The homology service will enable a user to rapidly query all proteins in all genomes in the current KBase reference database for a single protein sequence, or a set of protein sequences. Blast results will be returned at high speed. We will make the homology search functionality available in two places: (i) the search page on KBase.us, accessible without a KBase account; and (ii) as a method in the KBase narrative.

On the search page, users will be able to paste in one or more protein sequences either as plain text (for a single protein sequence) or FASTA format (for one or more protein sequences). The search will run on the fly, meaning only “fast” blast algorithms will be usable on the search page. The results will appear on screen as a table of homologous genes found for each query sequence. Users will be able to select all the results or a portion of the results for: (i) download in FASTA or TSV format, or (ii) copy into an existing narrative as a homology search object.

In the narrative, the user can upload a fasta file as a feature set, or they can search for features in the search page and copy to a narrative as a feature set. They can then run the “comprehensive homology search method”, which takes either a feature set as input, or has a text box where the user can copy and paste query sequences, as they did on the search page. They can select a number of optional features which allow them to specify cut-offs for returned homologs, sequence search algorithms, NRs to search against etc. The homologs are returns on screen with a variety of views, including tables and alignments. Data will be stored in the workspace as a FeatureHomologs object.

We will then add methods to the narrative that enable the user to select subsets of the homologs, to view the distribution of functions associated with the homologs, and to view and compare the functional context of all homologs.

**Longer description:**

This functionality is designed to allow a user to query all the reference genomes in KBase (or any other particular NR we wish to enable) for a single or set of query protein sequences. The system returns lists of homologs at high speed, along with homology values.

This is immediately useful to users who simply want to use KBase to get a list of homologs for a gene of interest.

However, this is really tailored as a feature to support the exploration of protein function. To better serve this second mission, we will add support methods in the narratives to view the functional distribution of homologous genes, to view the phylogenetic distribution of homologous genes, and to view and compare the contexts of gene homologs. The mock narrative demonstrates this.

The idea is to use the homology hits either directly to understand the function of a protein (by looking at the functions annotated to the homologs), or to use trends in the context of the homologous genes to identify function (by looking at functions of genes that are co-localized with all homologous genes).

Ultimately, I envision the output of these functions being piped into a multi-sequence analysis workflow. OR I envision homolog sets being queried against a set of user genomes sitting in a workspace using the user-data homology search workflow developed by Ranjan (with some refinements and improvements which we would do in the context of building the homology service).

I propose to build this now because it is among the most commonly requested features by our user base, and this sort of analysis is one of the first ways biologists make their way into our system. It has the potential to pull users into KBase, where they can begin to explore some of the numerous other functionalities we offer. Further, the output of this method feeds into numerous valuable other functions, including model gapfilling, multi-sequence analysis, co-expression analysis, and variation analysis.

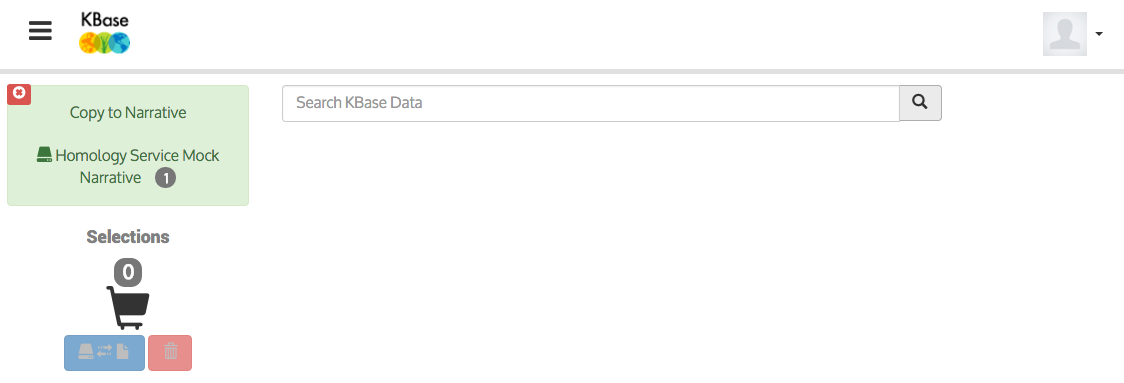
**Short user stories:**

1. User Gyorgy comes to the KBase home page. He queries on the main search page for “thermotoga cellulose”. He selects one of the genes that comes up, opening up the gene page and copying the protein sequence. He then selects “sequence search” and pastes the protein sequence. He gets a list of homologous genes, which he copies to a narrative and then exports and downloads as a fasta file. He uses the homologs to attempt to find a version of a gene he can express in E. coli for protein crystallization.
2. User Chris runs a metabolic model and finds a dead end pathway consisting of ten steps, only one of which is present in his genome. He selects the gene in the model viewer and exports it to his narrative as a feature set, which he labels “orphan protein”. He then selects the homology search method, enters in his “orphan protein” object, hits submit. The results pop up. He looks in the functional distribution and finds that many of the homologs are annotated with a different role than the original orphan protein. Furthermore, that role is associated with a gapfilled reaction in his model. Chris then looks at the functional context analysis, which shows that his protein has a context that is much more consistent with this gapfilled function than with the original function. Chris reannotates the protein and rebuilds the model.

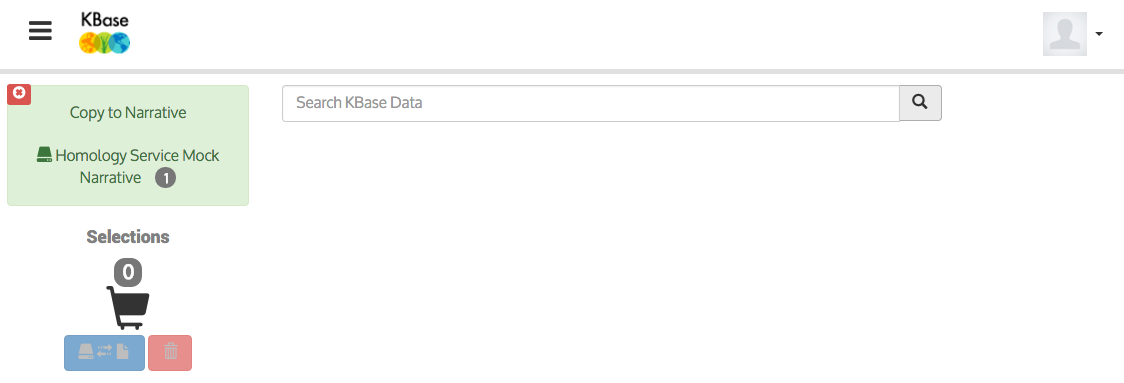
**Narrative mockup:**

https://narrative-ci.kbase.us/narrative/ws.3594.obj.1

**Search page mockup:**



protein sequence search



MRLNLFVKGMALGVMSLFLLIPAAFAAGEDTDTNVKPNEYQEKKLDINSSILRDQTKYEQSKKTSTVKTDIHFAEPPKTPGGSLSPQLFSDLKENPPKTSAHMMEEMNISFRDAAVSTPSGEDQEKQSSALLPIIFGCFIVFGIVVMFILIPKVNVKAEKK

Search kbase plants

Search kbase microbes

**Citation of code to be incorporated:**

?