Phylogenetic Clustering of Hexokinase 1 Paralogs Using String Kernels to produce Sequence Based Metric Distances

Valerie Chau and Jonas Grove

Introduction

Proteins are responsible for the majority of processes in the body. Protein function, a notoriously elusive challenge in bioinformatics, is determined, in part, by its structural shape, which is ultimately based on its amino acid sequence. Similarities in protein sequences indicate shared homology and function, allowing proteins to become a holistic source of knowledge.

A distance between two sequences is a comparable metric value, that can be further used to cluster organisms based on relative similarity. For a distance to be considered metric, the distance calculated must be symmetric, positive, satisfy the triangular inequality, and the distance must be non-existent between itself. This is especially vital in further clustering applications because close proximity to a common point, does not necessarily guarantee the same close relation between each other.

Comparing protein sequences traditionally involves aligning sequences via dynamic programming algorithms like Needleman-Wunsch (global alignment) or Smith-Waterman (local alignment). These methods can generate a score, but the score is not a metric, as it fails to satisfy the triangular inequality, and thus cannot be used to compare and cluster multiple sequences. Heuristics methods, like BLAST, compare a query sequence to a collective database to find similar sequences that come with a statistical significance score. However, BLAST doesn't generate a similarity score between sequences. Rather, BLAST merely detects similarity. Pairwise alignment methods are limited in that they do not define a metric in sequence space, meaning that those values cannot be compared between sequences for further clustering.

Thus, within the past decade, efforts to find an alternate method to pairwise alignments have been made. Smale and colleagues, in 2012, published an article titled, "Towards a Mathematical Foundation of Immunology and Amino Acid Chains", which described a mathematical model of generating metric distances between two sequences.**(1)** Smale describes using 'string kernels', which are functions that create the dot product from two sequences. To make the distance between *two* sequences as exact as possible, the distance between two amino acids in a sequence space, much like the distance between two physical points in space, is the shortest and most exact approximation of a distance. The distance between

two amino acids being compared is based on a substitution matrix, like BLOSUM which take into account chemical and physical properties. Smale's method includes the mathematical construction of kernels built upon previous kernels and takes in different inputs.

Kernel 1 (K_1) works on the space of pairs of amino acids, and because it directly compares two amino acids via substitution matrix, it is a metric measurement. Kernel 2 (K_2) is a kernel that works on a space of strings with the same length. Kernel 3 (K_3) works on a space of strings with different lengths. Both K_2 and K_3 both generate a dot product between every possible contiguous subsequence (*k*-mer, where *k* is the number of amino acids in the subsequence) between both two sequences by calling upon K_1 and summing the return values. Therefore because K_1 returns a metric measurement and the sum of a kernel is a kernel, both K_2 and K_3 can be considered a kernel that returns a metric measurement as well.

Using an implementation of the method described by Smale to calculate a metric distance between protein sequences, it is recognized that this method could be used to evaluate the similarities of paralogous proteins. In this procedure the enzyme Hexokinase 1 was chosen as a subject, as this enzyme catalyzes a key step in the glycolytic pathway by conversion of glucose to the activated molecule glucose-6-phosphate.**(2)** Because HK1 is a key component of metabolism, it is found in nearly all types of organisms and therefore was chosen as a test subject to investigate the ability of the implemented algorithm to distinguish differences between proteins, and to phylogenetically cluster the protein sequences together based on the distances generated from the described implementation of Smale's method.

It has been shown in work conducted by M. Cardenas that HK1 has the closest sequence similarities to organisms belonging to the same kingdom, and therefore it is hypothesized that the algorithm will cluster HK1 sequences in an analogous pattern.**(3)** For example, it is expected that if the implemented algorithm is operating as intended, the mammal, plant and single celled organism sequences are expected to form individual clusters.

Methods

Overview of distance methods

Using Smale and Colleagues (2012) and Bojoomi and Koehl (2017)'s previous work on string kernels as the theoretical base, our program calculates the distance between strings of amino acids. Our implementation of Smale's method reads in FASTA files containing the sequences, a blossom matrix text file, and asks for a user input beta value, used to scale the Blosum matrix. The program returns the distance between two protein sequences, normalized by length.

Much like Smale's method describes, the program constructs kernels based on the length of the input sequences. The function, Kernel 1 (K1), returns the value of a direct substitution of two amino acids taken from the beta-scaled amino acid substitution matrix. The function, Kernel 2 (K2), creates a dot product for every *k*-mer possible between sequences of the same length, where the function K1 is called upon, with the individual amino acids compared as input. The values returned by K1 are then summed together to generate the distance between the entire sequence. The function, Kernel 3 (K3), calculates the dot-product for every *k*-mer possible, for pairs of sequences with different lengths, where the shorter sequence's length is the limiting *k*-mer size. Much like the function K2, K3 calls upon K1 to return the direct 1:1 amino acid comparison and sums the results to compute the distance between the entirety of both sequences.

While K2 and K3 theoretically differ only in the length of the input sequences, they are fundamentally the same in process, and because both K2 and K3 rely on K1 to return the direct substitution value, *our* program is implemented in a hierarchical manner such that the kernels are constructed such that K3 calls on K2 and K2 calls on K1. K3 creates the *k*-mers to be inputted into K2. K2 iterates through the *k*-mer to call K1 for each amino acid comparison, and K1 returns the direct substitution value. K2 multiplies the returned values from K1, and K3 sums the values returned from K2. In our program, the function donkeyKong, converts the normalized K3 values into a distance.

Run time

As all *k*-mers of the shortest sequence length are created and compared, and because the normalization function calls for the K^3 value to be calculated thrice, the program is bound by an $O(n^4)$ runtime. While comparing two sequences of length 68-94 base pairs took less than a minute to complete. The hexokinase sequences selected varied in size from 500 - 950 bp produced a high computational cost.

Therefore, to reduce the run time, and increase the range of sequences that can be accepted, the program was adjusted such that $k = 10$ was the maximum k-mer size created.

While limiting the k-mer value was not a feature Smale and his colleagues implemented in the original 2012 paper, they tested their program with "sequences of length 9-37 amino acids", and thus most likely did not need to implement limits to the length of the sequence.**(1)** However, Bojoomi and Koehl (2017) in the article titled, "String kernels for protein sequence comparisons: improved fold recognition", used a "maximum length of the k-mers […] set to a small number for computational considerations." Testing a range of different maximum k-mer values, they found that "any value of kmax is possible, pending that the proper value for β is chosen".**(4)** As the actual distances have little meaning, compared to the relative distances, so long as the relationships between the sequences remain, reducing the k-mer size to a maximum of 10 should have no effect on the final clustering pattern. This was additionally tested by running the same 3 sample sequences of length 68-94 with the new k-mer limit, and verified by the program returning the same distance relationships between the three sequences.

Thus, to test the 10 hexokinase sequences, all ranging in sequence length of 500-950 base pairs, it was imperative that the maximum length be set such that the information could be processed within a reasonable time frame. Establishing a maximum of 100 and 50 base pairs all yielded a high computational cost, and a maximum value of 10 base pairs (such that a 10-mer is the longest subsequence being compared) was established for comparing sequences larger than 9 bp.

Clustering methods

If only two to three sequences are being compared using the above algorithm, it is trivial to analyze the distances generated and interpret the relationships between the different organisms. However, if multiple sequences from different species are being compared, the relationships between these sequences becomes difficult to interpret and thus it becomes important to develop a way of efficiently interpreting these results.

This requirement motivated that the results generated from the Smale implementation be stored in a graph, represented as an adjacency matrix, in which every node is a protein sequence and every edge represents the distance between the two connected organisms. This graph was then used to generate a minimum spanning tree, using an implementation of Kruskal's algorithm derived from an online source**(5)**, and the maximum K nodes were removed to generate K-1 clusters.**(6)**

In order to test the efficacy of the implementation described above, the Hexokinase 1 sequence endogenous to nine different organisms was used as a test input. The input sequences were sourced from the NCBI database and the organisms were chosen such that there would be representatives from three different groups including; vertebrates, plants, and single celled organisms.

It was also of interest to see what organisms would cluster when the number of declared clusters to be formed was increased or decreased. For this reason, the clustering algorithm was run using the organism distance graph and different K values as inputs.

Results

Distance Results

The results, displayed in Table 1, show the distances returned when comparing the sequence in the corresponding row and column. For each row, the darkness in color indicates the relative shortness in distance of the two corresponding hexokinase sequences between other species in the same row. Only the shortest five distances were indicated with a color. All other distances were left unmarked. The diagonal, where each sequence was compared to its own species, showed a unanimous distance = 0.

Table 1: Table with distances between HK1 sequences from organisms listed, with darkness of color denoting relative closeness between sequences for organisms in row.

	Human	Orangutan	House Mouse	Brown Rat	Zebra Fish	A. Thaliana	*Z.* mays	Bacteroidetes	Yeast
Human	**0**	0.001703562507	0.005540196414	0.003509587464	0.006556515147	0.02125364837	0.024038478 7	0.019332981 16	0.021013 96427
Orangutan	0.001703562507	**0**	0.004843471958	0.003250906107	0.006090300 69	0.020832419 74	0.023123277 31	0.019042106 72	0.020097 58335
House Mouse	0.005540196414	0.004843471958	**0**	0.004599556135	0.006082008838	0.020826622 07	0.023486276 74	0.020222472 07	0.019422 20873
Brown Rat	0.003509587464	0.003250906107	0.004599556135	**0**	0.006116903778	0.020826622 07	0.024023151 52	0.018248075 98	0.019855 76633
Zebra Fish	0.006556515147	0.006090300 69	0.006082008838	0.006116903778	**0**	0.021821293 34	0.023503139 12	0.021050723 77	0.020617 68359
A. thaliana	0.02125364837	0.020832419 74	0.020823209 59	0.020826622 07	0.021821293 34	**0**	0.017639988 23	0.024345967 74	0.025320 79575
Z. mays	0.024038478 7	0.023123277 31	0.023486276 74	0.024023151 52	0.023503139 12	0.017639988 23	**0**	0.027679829 69	0.023697 79993
Bacteroidetes	0.019332981 16	0.019042106 72	0.020222472 07	0.018248075 98	0.021050723 77	0.024345967 74	0.027679829 69	**0**	0.016638 64606
Yeast	0.021013 96427	0.020097 58335	0.019422 20873	0.019855 76633	0.020617 68359	0.025320 79575	0.023697 79993	0.016638 64606	**0**

Key:

	Smallest Distance
	2nd Smallest Distance
	3rd Smallest Distance
	4th Smallest Distance
	5th Smallest Distance

Clustering results

The clustering algorithm used on the adjacency matrix generated groups consistent with the phylogenetic classifications of the organisms from which the sequences were derived, as shown in Figure 1A. The mammals, plants and microorganisms all grouped into individual clusters, when the cluster number was set to 3 (E=2), and upon increasing the E value the clusters which resulted were observed.

When the organisms were clustered into 4 groups, the plant species *A. thaliana* and *Z. mays* split into two individual groups, as shown in Figure 1B. The observation of this result suggests that the protein sequences endogenous to the two plants used in the procedure are more distant from each other than the sequences of fish and mammals, and fungi and bacteria. When the organisms were clustered into five groups, the fungal and bacterial HK1 sequences split into two individual groups, as shown in Figure 1C, thus suggesting that the bacteria species and the fungi (yeast) sequences are more distant from each other than the vertebrate sequences are distant from each other.

Upon further increase of the number of clusters to 6 (E=5), the fish separated into its own cluster, as would be expected based on its evolutionary distance from mammals. Increasing the number of clusters to 7 (E=6) resulted in the mouse sequence forming its own individual cluster, which interestingly suggests that the rat sequence is more similar to the two primate sequences than the mouse sequence. When the cluster number was set to 8 (E=7), the only organisms that remained clustered together were the Human and Orangutan, which is consistent with those species being the most closely related.

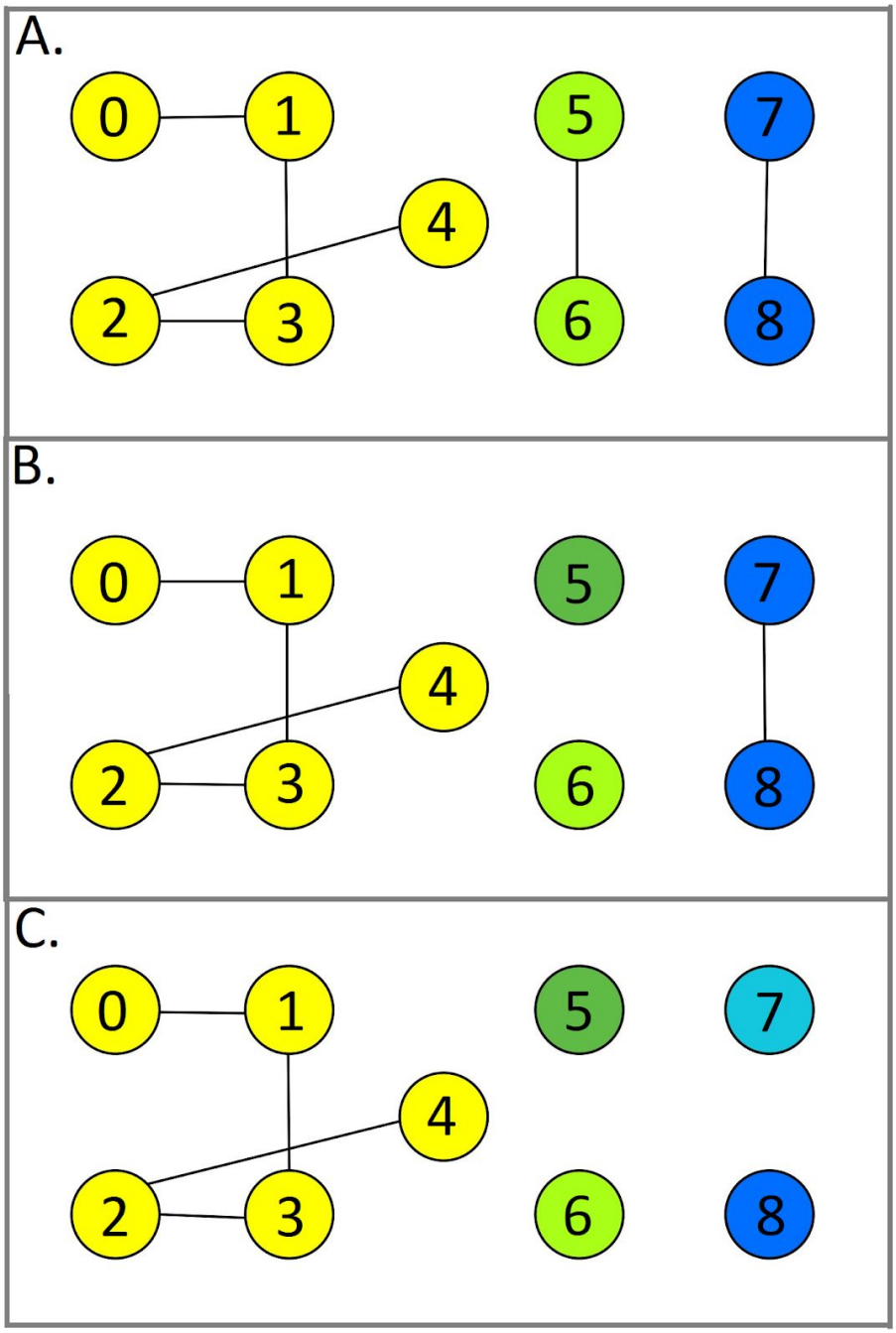

Figure 1. Clustering Results of Nine Hexokinase Sequences. Figure 1 shows the clustering patterns using different K values and Hexokinase 1 sequences from nine organisms in the following order; [0:Human,1:Orangutan,2:Mouse,3:Rat,4:Fish,5:Arabidopsis,6:Corn,7:Bacteroidetes,8:Yeast]. Figures 1A, 1B, and 1C show the results generated from the cluster numbers being set to 3, 4, and 5 respectively

Discussion

Discussion on Clustering

The initial result observed when the cluster number was set to 3 is consistent with what would be expected based on the previous annotation of HK1 sequence divergence, however the results were not entirely consistent with an evolutionary perspective.**(7)** One instance of this inconsistency is that the yeast and bacteroidetes species formed an independent group, despite the fact that yeast are more closely related evolutionarily to mammals than to bacteria, as shown in figure 2.**(7)** This observation suggests that the HK1 sequence endogenous to yeast was conserved as bacteria evolved into the eukaryotic species, yeast.

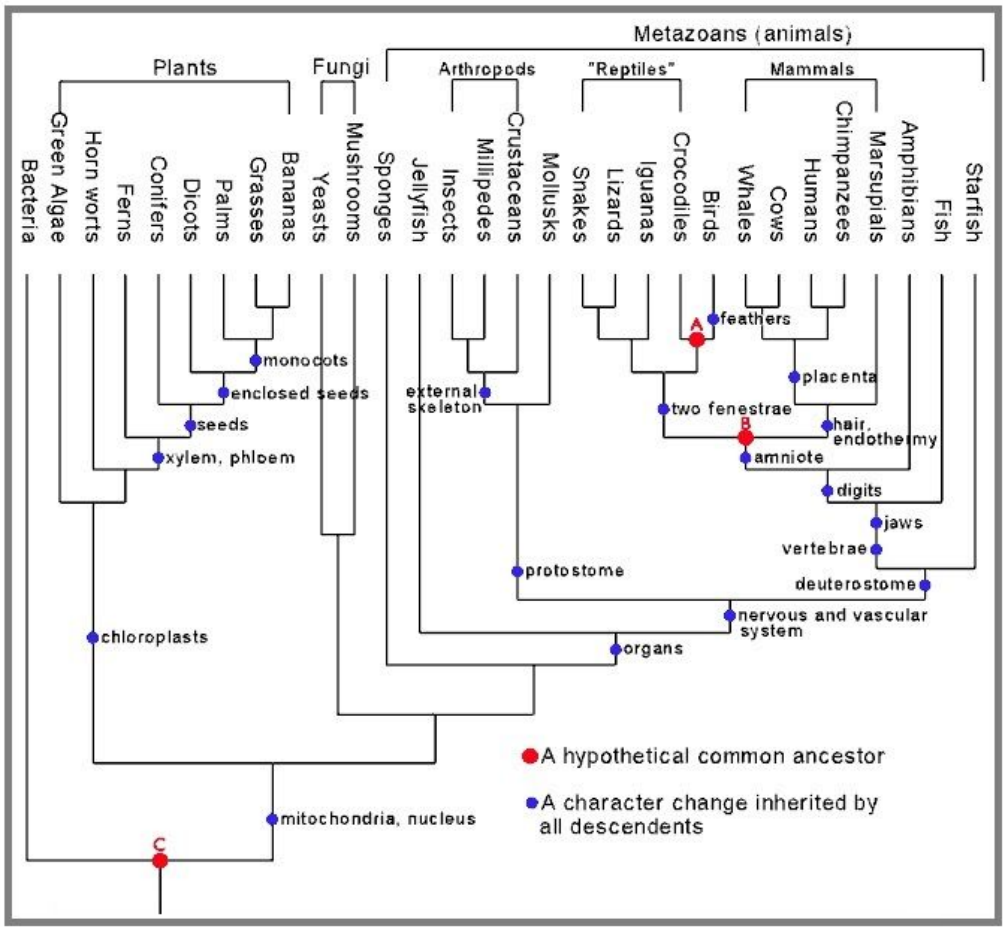

Figure 2. Tree of Life. Figure X shows the hierarchical phylogenetic relationships between different groups of organisms over the course of evolution.**(8)**

Another peculiar observation is that when the cluster number was set to four, the two plant species formed individual clusters, while it would be expected from an evolutionary perspective that the

fish would separate from the mammals due to the evolutionary distance between two plants being less than the evolutionary distance between a fish and a mammal**(7)**.

Another interesting observation is that when the cluster number was set to seven, the mouse formed its own individual cluster, rather than the rat and mouse forming one individual cluster, as would be expected from an evolutionary perspective. This suggests that the HK1 sequence endogenous to rats is more similar to the orangutan HK1 sequence than it is to its more closely related species, the mouse.**(7)**

It is possible that the reason for these anomalies is due to the HK1 sequences evolving at a different time scale than speciation occurred, however it should also be acknowledged that these discrepancies could be due to the sourcing of the sequences themselves. The sequences used in this procedure were sourced from different database entries on NCBI, which could have led to peculiarities in the clustering analysis. A more robust method to perform this type of comparison could be to source all of the sequence information from the same research group, which would eliminate inconsistencies that might result from different sampling practices.

Future Adjustments

While the program is able to calculate the expected distances, one future adjustment that can be implemented is to increase usability, will be to modify the program such that the run-time does not require limiting the kernel size to 10 base pairs. A more thorough analysis of the sequences being compared would involve completing the K^3 distance to the intended full length. Additionally, Nojoomi and Koehl (2017), implemented different pairs of ß-value and sizes K-max, but "suggest(ed) using the pair (β, kmax)=(0.2,10)".**(7)** Increasing the beta-value to 0.2 might allow for an increase in accuracy of our results. In terms of clarifying the distance between the chosen hexokinase sequences, one alternative method that could be implemented is using different BLOSUM substitution matrices. While Blosum62 was used to compare the hexokinase sequences in this specific study, additional trials, where Blosum80 and Blosum45 are interchanged between species within the same kingdoms could change the way the organisms cluster, yielding groups that align more towards what was shown in the **Fig 2**. One possible limitation of using the Smale method to calculate the distance between proteins is that it does not capture conserved domains between different proteins. This could be limiting, because two proteins could have very similar catalytic sites, but differ significantly in sequence and/or size. The Smale method would calculate the distance between these sequences to be large, however in reality the proteins may be very similar in catalytic function. With a more accurate clustering algorithm, one possible further improvement

would be to additionally look for similar domains, which could provide information as to which regions of the sequence are conserved, if any. This information could be used to detect proteins from different organisms which are conserved in function, but differ more significantly in sequence or size.

Resources

(1) Shen, W.-J.; Wong, H.-S.; Xiao, Q.-W.; Guo, X.; Smale, S. Towards a Mathematical Foundation of Immunology and Amino Acid Chains. *arXiv:1205.6031 [cs, q-bio, stat]* 2012.

(2) Boyer, R. Chapter 15.1, The Energy Metabolism of Glucose. *Concepts in Biochemistry, 3E-2006, John Wiley and Sons.*

(3) Cárdenas, M. L.; Cornish-Bowden, A.; Ureta, T. Evolution and Regulatory Role of the Hexokinases. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1998, *1401* (3), 242–264. https://doi.org/10.1016/S0167-4889(97)00150-X.

(4) Nojoomi, S.; Koehl, P. String Kernels for Protein Sequence Comparisons: Improved Fold Recognition. *BMC Bioinformatics* 2017, *18* (1), 137. https://doi.org/10.1186/s12859-017-1560-9.

(5) Kruskal's Minimum Spanning Tree Algorithm | Greedy Algo-2. GeeksforGeeks, 2012.

(6) Kleinberg, M.; Tardus, E. Chapter 4.7, Clustering. *Algorithm Design - Cornell University-2006*, *Pearson Education.*

(7) Letunic, I.; Bork, P. Interactive Tree Of Life (ITOL): An Online Tool for Phylogenetic Tree Display and Annotation. *Bioinformatics* 2007, *23* (1), 127–128. https://doi.org/10.1093/bioinformatics/btl529.

(8) 29+ Evidence for Macroevolution: Phylogenetics http://www.talkorigins.org/faqs/comdesc/phylo.html (accessed Mar 12, 2020).

Sequence Resources

(1) hexokinase 1 [Homo sapiens] - Protein - NCBI https://www.ncbi.nlm.nih.gov/protein/AAA52646.1 (accessed Mar 13, 2020).

(2) hexokinase-1 [Pongo abelii] - Protein - NCBI https://www.ncbi.nlm.nih.gov/protein/NP_001125344.1 (accessed Mar 13, 2020).

(3) hexokinase [Mus musculus] - Protein - NCBI https://www.ncbi.nlm.nih.gov/protein/AAB57759.1 (accessed Mar 13, 2020).

(4) hexokinase-1 [Rattus norvegicus] - Protein - NCBI https://www.ncbi.nlm.nih.gov/protein/NP_036866.1 (accessed Mar 13, 2020).

(5) Hexokinase 1 [Danio rerio] - Protein - NCBI https://www.ncbi.nlm.nih.gov/protein/AAH67330.1 (accessed Mar 13, 2020).

(6) hexokinase 1 [Arabidopsis thaliana] - Protein - NCBI https://www.ncbi.nlm.nih.gov/protein/AEE85590.1 (accessed Mar 13, 2020).

(7) hexokinase-1 [Zea mays] - Protein - NCBI https://www.ncbi.nlm.nih.gov/protein/ACG47843.1 (accessed Mar 13, 2020).

(8) hexokinase [Bacteroides thetaiotaomicron] - Protein - NCBI https://www.ncbi.nlm.nih.gov/protein/KAB4438887.1 (accessed Mar 13, 2020).

(9) hexokinase 1 [Saccharomyces cerevisiae S288C] - Protein - NCBI https://www.ncbi.nlm.nih.gov/protein/NP_116711.3 (accessed Mar 13, 2020).