

Manual for HGT Calculator

- **Install HGT Calculator:**

Download HGT-Calculator.zip and unpack it. Save HGT-Calculator.jar and the file folder HGT-Calculator_lib at the same location of your workspace.

- **Start HGT Calculator:**

Double click on HGT-Calculator.jar or execute the file from the terminal via

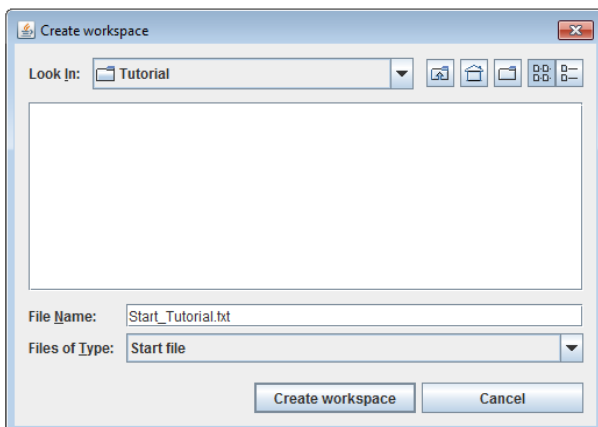
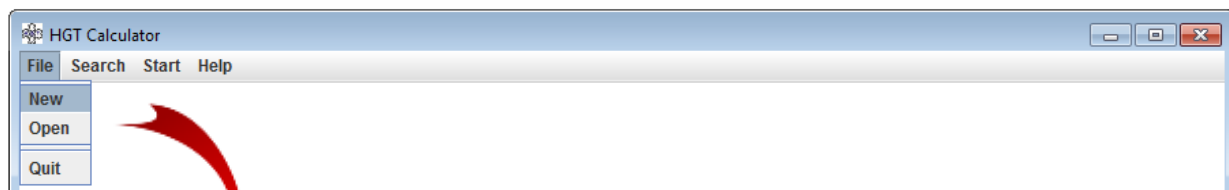
```
java -jar HGT-Calculator.jar
```

Please note: Java applications can sometimes generate problems with antivirus software and permissions. If so, add HGT-Calculator.jar to the whitelist of your antivirus program and change HGT-Calculator.jar to be an executable jar file! On Windows you can also use the batch file Start_HGT-Calculator.bat to start HGT Calculator.

Other problems can be caused by updates of the UniProt database. This can be fixed by downloading the newest version of the UniProt JAPI ([uniprot-japi-client.zip](#)) and copying the unpacked JARs into “HGT-Calculator_lib”. Be careful to keep clustalw2.exe and iprscan5_lwp.pl in HGT-Calculator_lib to start this external tools inside HGT Calculator.

- **File > New:** Choose a directory for creating a new workspace. The default is the directory containing the HGT-Calculator.jar. Enter the name of the new start file and click on “Create workspace”. Do not select an existing text file! If you want to use an existing start file, copy the text into the new start file of the workspace.

Please note: The name of the start file needs to start with “Start_” and ends with “.txt” !



→ Creates workspace architecture:

```
workspace\Data\Download \Blast
                                \DownloadLists
                                \Fasta
                                \Genes
                                \Input
                                \Log
                                \Newick
                                \Output
                                \UniProt
```

→ Creates a new start file, where you need to add your information:

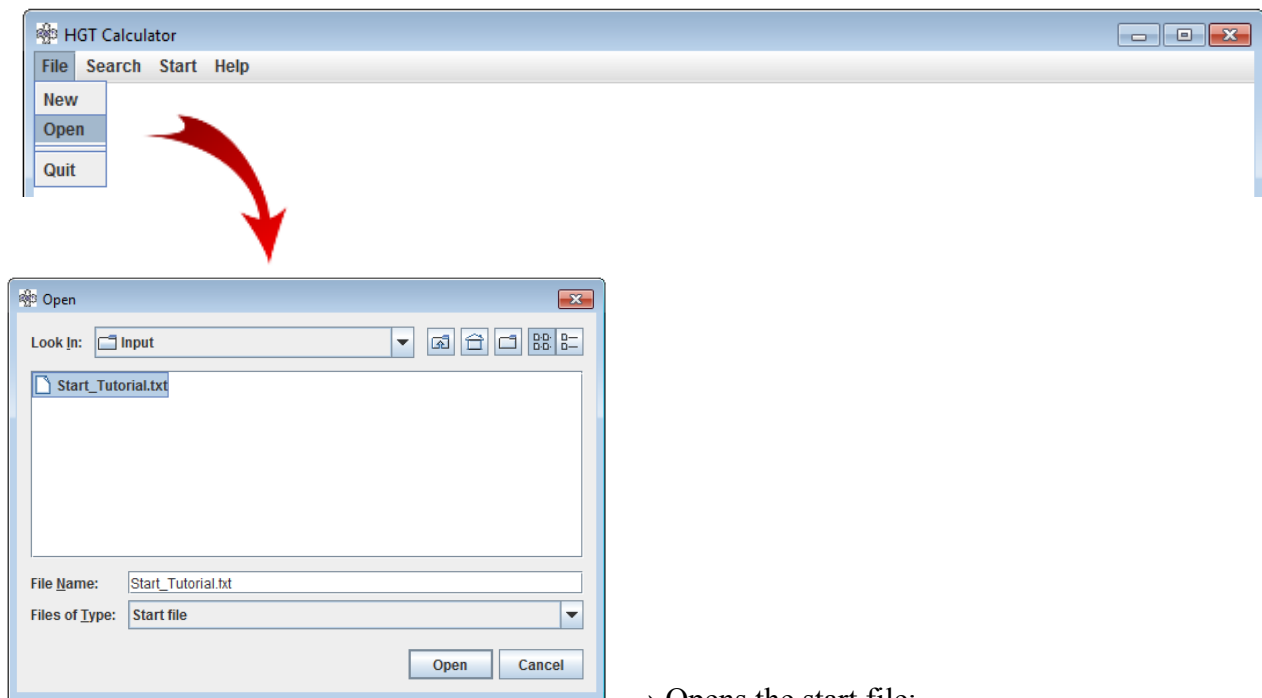
Please note: At the moment, HGT Calculator is only tested for the detection of HGT from bacteria into protozoa!

Name of the data set > Tutorial
Transfer from > Bacteria
Transfer into > Protozoa
Protein abbreviations >
Prokaryotic organisms >
Eukaryotic organisms >

Name of the data set > Tutorial
Transfer from > Bacteria
Transfer into > Protozoa
Protein abbreviations > **ICDH**
Prokaryotic organisms > **Streptomyces, ...**
Eukaryotic organisms > **Tetrahymena, ...**

If the start file template is not opened automatically, because on your computer .txt is not linked with an editor, you have to open it manually and add your information, before you can open this workspace.

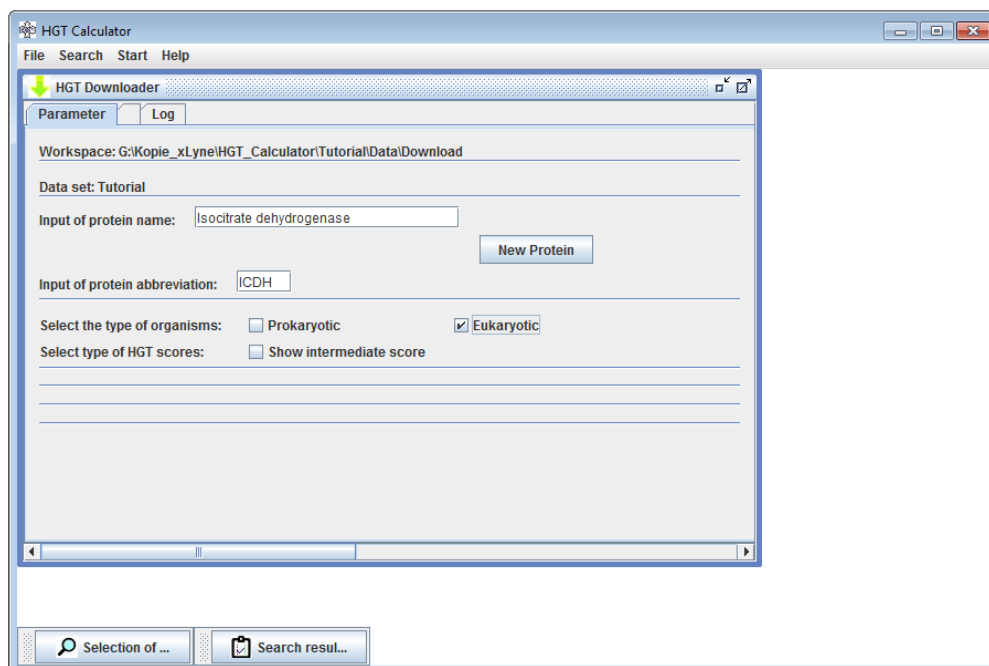
- **File > Open:** Choose a start file from the Input directory for opening a workspace.



→ Opens the start file:

Name of the data set > Tutorial
Transfer from > Bacteria
Transfer into > Protozoa
Protein abbreviations > ICDH
Prokaryotic organisms > Streptomyces, Micrococcus, Aquifex, Cytophaga, Chlorobium, ...
Eukaryotic organisms > Tetrahymena, Paramecium, Leishmania, Trypanosoma, Cryptosporidium, ...

→ Opens HGT Downloader:

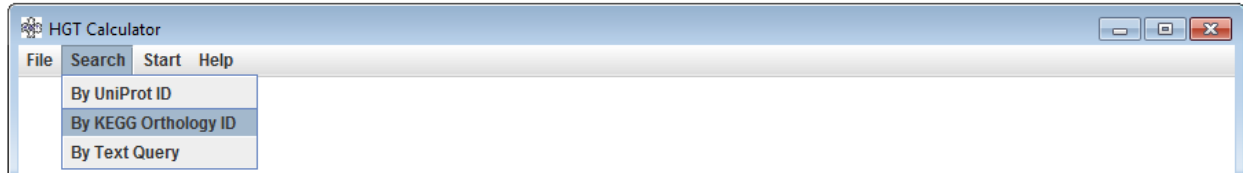


→ Enter a protein name (e.g. “Isocitrate dehydrogenase”).

→ Enter a protein abbreviation (between two and four letters, e.g. “ICDH”).

→ Check “Eukaryotic” to start with the search for eukaryotic sequences.

Search options of HGT Downloader:



- **Search > By UniProt ID:**

Search in the UniProt database with a known UniProtID as query:

Input of UniProt ID from one organism:

- Enter an UniProtID
- Gets corresponding KEGG ID
- Gets corresponding KEGG Orthology ID
- Searches for other entries with this KEGG Orthology ID for selected organisms

- **Search > By KEGG Orthology ID:**

Search in KEGG database with a known KEGG Orthology ID:

Input of KEGG Orthology ID for Isocitrate dehydrogenase:

- Enter a KEGG Orthology ID or “get” to search for the KEGG Orthology ID of the protein name
- Searches for other entries with this KEGG Orthology ID for selected organisms

- **Search > By Text Query:**

Search in the UniProt database with organism + protein name as query

Input of Query organism:

- Enter an organism or “all” (search for all selected organisms)
- If text based search failed enter a new query:
 - other protein name (e.g. “Arabidopsis > peptidase M1”, instead of “Aminopeptidase”)
 - or InterPro ID (e.g. “Leishmania > IPR004439”)

Organism	KEGG ID	Gene name	UniProt ID	Selected	InterPro	KEGG Orthology ID
Leishmania major	lma.LMJF_33_2550	(RefSeq) putative isocitrate dehydrogenase [...]	Q4Q3T2	<input checked="" type="checkbox"/>	IPR001804 Isocitrate/isopropylmalate_DH; IPR0044...	K00031
Plasmodium chabaudi	pcb.PC000375.04.0	(RefSeq) hypothetical protein	Q4XPY6	<input type="checkbox"/>	IPR004790 Isocitrate_DH_NADP; IPR024084 IsoPro...	K00031
Plasmodium chabaudi	pcb.PC000412.00.0	(RefSeq) isocitrate dehydrogenase (NADP), ...	Q4Y7E4	<input type="checkbox"/>	IPR004790 Isocitrate_DH_NADP; IPR019818 IsoCih...	K00031
Plasmodium chabaudi	pcb.PC302184.00.0	(RefSeq) hypothetical protein	Q4X8E0	<input type="checkbox"/>	IPR004790 Isocitrate_DH_NADP; IPR024084 IsoPro...	K00031
Plasmodium falciparum	pfa.PF13_0242	(RefSeq) isocitrate dehydrogenase (NADP), ...	Q8I6T2	<input checked="" type="checkbox"/>	IPR004790 Isocitrate_DH_NADP; IPR019818 IsoCih...	K00031
Paramecium tetraurelia	ptm.GSPATT00005218001	(RefSeq) hypothetical protein	A0BPK2	<input checked="" type="checkbox"/>	IPR004790 Isocitrate_DH_NADP; IPR019818 IsoCih...	K00031
Paramecium tetraurelia	ptm.GSPATT00009194001	(RefSeq) hypothetical protein	A0BPK2	<input type="checkbox"/>	IPR004790 Isocitrate_DH_NADP; IPR019818 IsoCih...	K00031
Paramecium tetraurelia	ptm.GSPATT00020509001	(RefSeq) hypothetical protein	A0DUY0	<input type="checkbox"/>	IPR004790 Isocitrate_DH_NADP; IPR019818 IsoCih...	K00031
Paramecium tetraurelia	ptm.GSPATT00022820001	(RefSeq) hypothetical protein	A0E2Y1	<input type="checkbox"/>	IPR004790 Isocitrate_DH_NADP; IPR019818 IsoCih...	K00031

Please note:

Additional search parameter: search only for prokaryotic or eukaryotic sequences

→ Check boxes “Prokaryotic” / ”Eukaryotic”

Selection of query organisms: → Check boxes are checked for all organisms from the start file.

→ Change of selected organisms is possible for each new search.

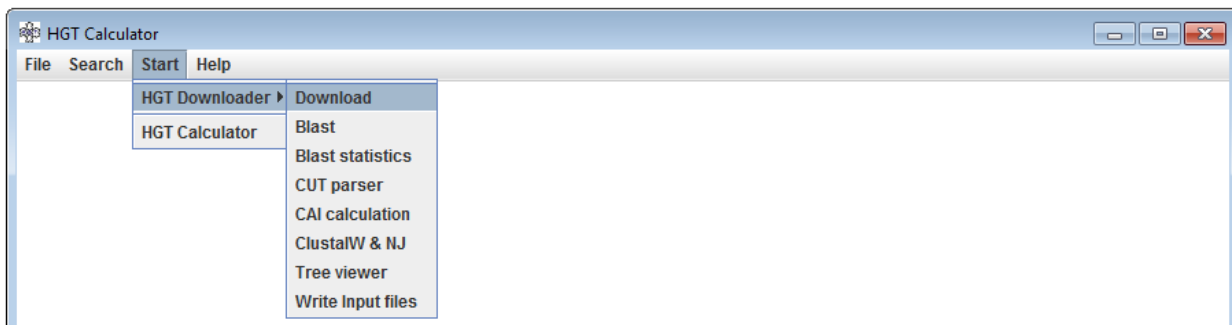
The screenshot shows a web-based interface titled "Selection of Organisms". It has two tabs: "Eukaryotes" (selected) and "Prokaryotes". Under the "Eukaryotes" tab, there are several categories of organisms with checkboxes for selection. The categories and their sub-items are: Ciliophora (Paramecium, Tetrahymena), Euglenozoa (Leishmania, Trypanosoma, Angomonas, Strigomonas), Apicomplexa (Babesia, Cryptosporidium, Plasmodium, Toxoplasma, Theileria), Amoebozoa (Dictyostelium, Entamoeba), Diatoms (Phaeodactylum, Thalassiosira), Other protozoa (Giardia, Naegleria, Ostreococcus, Phytophthora, Trichomonas), Fungi (Aspergillus, Neurospora, Penicillium, Candida, Saccharomyces, Schizosaccharomyces), Viridiplantae (Arabidopsis, Glycine, Oryza, Populus, Selaginella, Zea), and Metazoa (Caenorhabditis, Danio, Drosophila, Gallus, Homo, Xenopus).

The screenshot shows the same "Selection of Organisms" interface, but with the "Prokaryotes" tab selected. The categories and their sub-items are: Actinobacteria (Micrococcus, Streptomyces), Cyanobacteria (Cyanotheca, Nostoc, Synechococcus, Synechocystis), Alpha Proteobacteria (Brucella, Chlamydia, Rickettsia, Rhodospirillum, Sinorhizobium, Wolbachia), Other Proteobacteria (Desulfovibrio, Escherichia, Helicobacter, Neisseria, Burkholderia, Vibrio), Other bacteria (Aquifex, Bacillus, Chlorobium, Cytophaga, Deinococcus, Thermotoga), and Symbionts of protozoa (Candidatus Kinetoplastibacterium).

Begin a search always with the eukaryotic sequences:

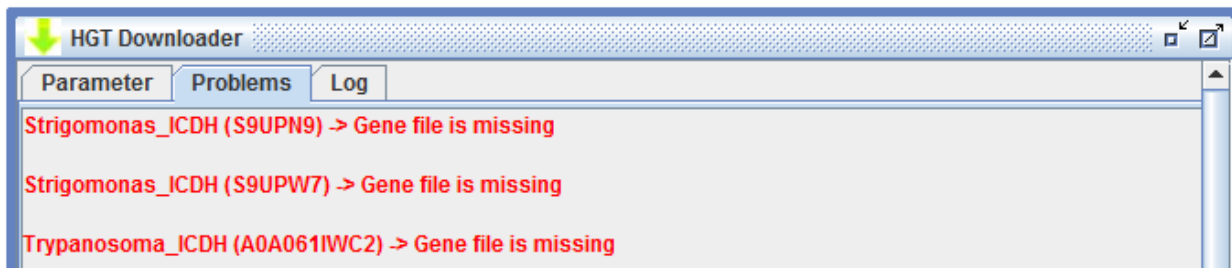
- Their subtypes are needed for clustering of the prokaryotic subtypes.
- Selected sequences are saved in two download lists for this protein (euk, prok) after clicking on the first cell of the search results table which contains the protein abbreviation and the mode of the search (e.g. “ICDH | prok”).
- If there are sequences without protein domains, InterProScan is used to add IPR numbers to the downloaded protein file later on.

For organisms which are not listed in KEGG (e.g. Strigomonas) you have to start an additional text search after the search with a KEGG ID to get all sequences!

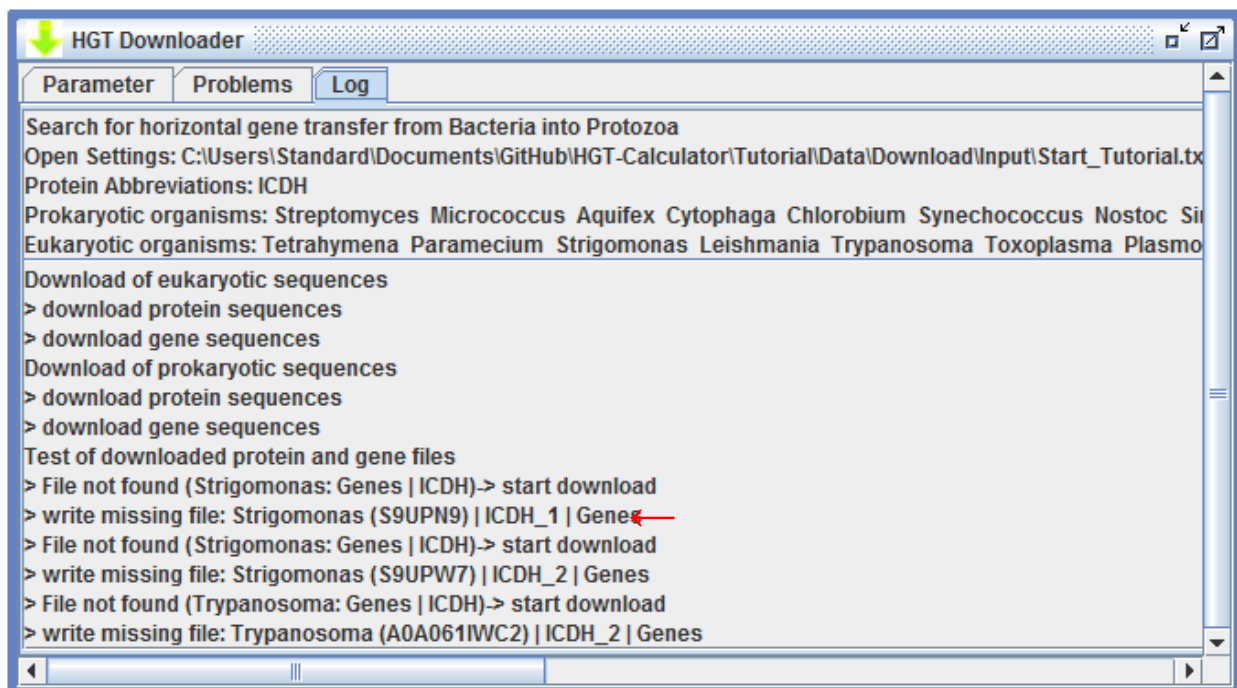


- **Start > HGT Downloader > Download:**

Starts download of all protein and nucleotide files of the download lists for this protein.



→ If there is no UniProtID or KEGG ID for the entry, the program shows an error on the tab “Problems”, which usually can be ignored. Such files are added by other accession numbers later on. Whether they were added properly or not you can check on the tab “Log”.

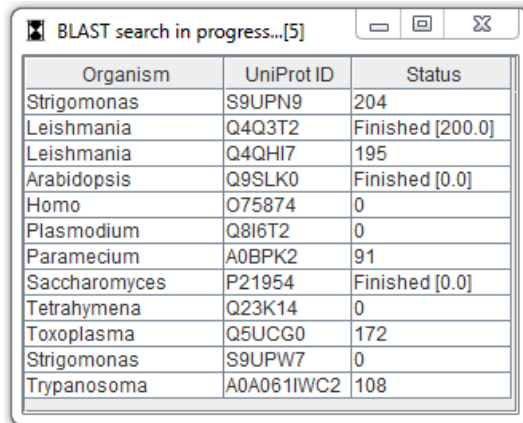


Please note: The complete log file is saved in \workspace\Data\Download\Log\Log_Tutorial\.

- **Start > HGT Downloader > BLAST:**

Starts BLAST search in the UniProt database for each entry of the download lists.

→ Saves the number of BLAST hits from Archaea, Bacteria and Eukaryotes, the Alien Index (AI) and a possible donor organism (if available) in the first line of the file.



Organism	UniProt ID	Status
Strigomonas	S9UPN9	204
Leishmania	Q4Q3T2	Finished [200.0]
Leishmania	Q4QHI7	195
Arabidopsis	Q9SLK0	Finished [0.0]
Homo	O75874	0
Plasmodium	Q8I6T2	0
Paramecium	A0BPK2	91
Saccharomyces	P21954	Finished [0.0]
Tetrahymena	Q23K14	0
Toxoplasma	Q5UCG0	172
Strigomonas	S9UPW7	0
Trypanosoma	A0A061IWC2	108

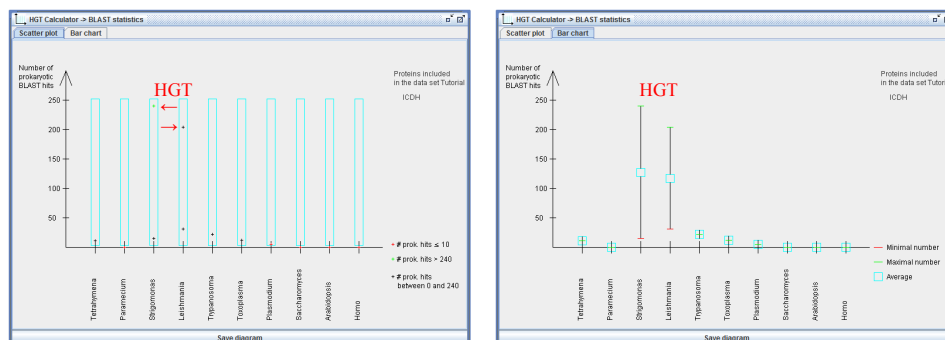
→ Alien Index of 200 → HGT

→ Alien Index of 0 → not HGT

- **Start > HGT Downloader > BLAST statistics:**

Draws two plots for visualizing the ratio of BLAST hits in this data set.

→ One is a bar diagram and the other one is a cross plot, which can be saved in \workspace\Data\Download\Blast\Blast_Tutorial_statistics_chart.jpg and \workspace\Data\Download\Blast\Blast_Tutorial_statistics_plot.jpg.



- **Start > HGT Downloader > CUT parser:**

Saves a codon usage table in EMBOSS format in a new file with Codon Usage Database format, which is needed for CAI calculation.

- **Start > HGT Downloader > CAI calculation:**

Downloads all needed Codon Usage Tables from the codon usage database (<http://www.kazusa.or.jp/codon>) and creates a file with average GC content in all selected organisms.

→ in workspace\Data\Download\Input\GC_Content_Tutorial.txt

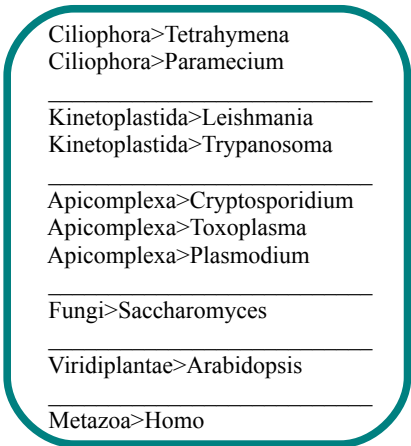
Tetrahymena	CodingGC_32.53	1stGC_38.64	2ndGC_31.25	3rdGC_27.69
Paramecium	CodingGC_31.71	1stGC_37.51	2ndGC_30.33	3rdGC_27.29
Leishmania	CodingGC_62.14	1stGC_61.94	2ndGC_49.76	3rdGC_74.73
Trypanosoma	CodingGC_50.73	1stGC_57.21	2ndGC_43.62	3rdGC_51.35
Cryptosporidium	CodingGC_33.28	1stGC_40.35	2ndGC_34.78	3rdGC_24.72
Toxoplasma	CodingGC_56.41	1stGC_60.03	2ndGC_47.52	3rdGC_61.68
Plasmodium	CodingGC_27.59	1stGC_37.98	2ndGC_27.93	3rdGC_16.85
Saccharomyces	CodingGC_39.77	1stGC_44.58	2ndGC_36.64	3rdGC_38.10
Arabidopsis	CodingGC_44.59	1stGC_50.84	2ndGC_40.54	3rdGC_42.38
Homo	CodingGC_52.27	1stGC_55.72	2ndGC_42.54	3rdGC_58.55

→ Calculates CAI values (Xia, 2007) and saves them in the download lists.

- **Start > HGT Downloader > ClustalW & NJ:**

Creates a file with taxonomic clusters of all selected organisms.

→ in workspace\Data\Download\Input\Cluster_Tutorial.txt



```
Ciliophora>Tetrahymena
Ciliophora>Paramecium

Kinetoplastida>Leishmania
Kinetoplastida>Trypanosoma

Apicomplexa>Cryptosporidium
Apicomplexa>Toxoplasma
Apicomplexa>Plasmodium

Fungi>Saccharomyces

Viridiplantae>Arabidopsis

Metazoa>Homo
```

Enter two download lists for creating files with all sequences needed for tree reconstruction in FASTA format:

Input of data sets for creating tree file and type of the sequences (- > protein):

→ e.g. “DownloadList_ICDH.txt; DownloadList_ICDH_prok.txt > protein” or
“DownloadList_ICDH.txt; DownloadList_ICDH_prok.txt > gene”
for protein or gene tree

→ Enter “- > protein” or “- > gene” to use the the default path.

→ Creates the input files for phylogenetic analysis in
workspace\Data\Download\Fasta\Fasta_Tutorial\.

→ Runs ClustalW2 (Larkin et al., 2007) for multiple sequence alignment and
creating Neighbor Joining trees.

→ The trees are saved in newick format

e.g. \Newick\Newick_Tutorial\ICDH_newickTree_prot.txt

\Newick\Newick_Tutorial\ICDH_newickTree_small_prot.txt

\Newick\Newick_Tutorial\ICDH_newickTree_single_Toxoplasma_prot.txt

→ You need three trees:

1. One large tree with all sequences (eukaryotic and prokaryotic)
2. One small tree with all eukaryotic sequences and just one prokaryotic sequence per eukaryotic protein subtype
3. A single tree for each organism with just the sequences of this single organism, a fungus, a plant, a metazoan, and one bacterial sequence per eukaryotic protein subtype

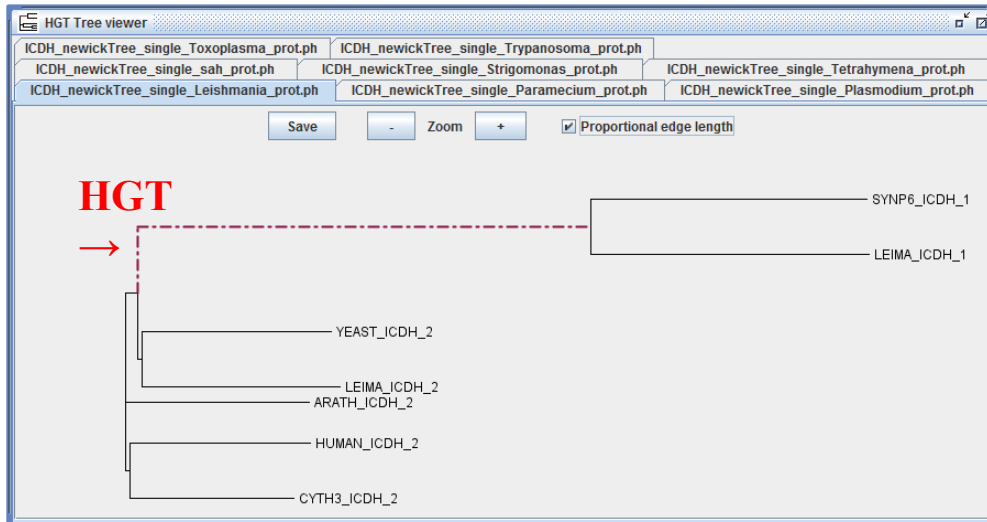
This single bacterial sequences are chosen randomly.

You can select any bacterium of the data set by entering the abbreviation of the bacterium after your path to the download lists.

(e.g. “- > protein # BACSU” for *Bacillus subtilis*)

- **Start > HGT Downloader > Tree viewer**

Shows all single trees to estimate the influence of the selected bacterium on the outcome of the phylogenetic analysis.



- **Start > HGT Downloader > Write Input files**

Writes the input files for HGT Calculator

→ Input_Tutorial_first.txt → lines ordered by protein

→ in workspace\Data\Download\Input\

Leishmania	prot>Q4QBD8	ADH_1	lma:LMJF_23_0360
Arabidopsis	prot>P06525	ADH_1	ath:AT1G77120
Homo	prot>P07327	ADH_1	hsa:124
Toxoplasma	prot>B9QJU6	ADH_1	no KEGG
Saccharomyces	prot>P00330	ADH_1	sce:YOL086C
Paramecium	prot>A0BN51	ADH_1	ptm:GSPATT00030606001
Tetrahymena	prot>A4VEL4	ADH_1	tet:THERM_00374939
Plasmodium	prot>Q4XZU7	ADH_1	pcb:PC103808.00.0
Trypanosoma	prot>Q4DY23	ADH_2	tcr:506357.50
Saccharomyces	prot>P10127	ADH_2	sce:YGL256W
Leishmania	prot>Q4Q784	ADH_2	lma:LMJF_30_2090
Homo	prot>Q8IWW8	ADH_2	hsa:137872
Cryptosporidium	prot>Q5CW43	ADH_2	cpv:cgd8_1700
Leishmania	prot>Q4Q3T2	ICDH_1	lma:LMJF_33_2550
Leishmania	prot>Q4QH17	ICDH_2	lma:LMJF_10_0290
Tetrahymena	prot>Q23W20	ICDH_2	tet:THERM_00794110
Paramecium	prot>A0BPK2	ICDH_2	ptm:GSPATT00005218001
Trypanosoma	prot>Q580Y9	ICDH_2	tbr:Tb927.8.3690
Toxoplasma	prot>Q5UCG0	ICDH_2	tgo:TGME49_066760

→ Input_Tutorial_second.txt → lines ordered by organism, with CAI values

Tetrahymena	prot>A4VEL4	ADH_1	tet:THERM_00374939	0.716
Tetrahymena	prot>Q23W20	ICDH_2	tet:THERM_00794110	0.711
Tetrahymena	prot>I7M6H8	IMP_D_1	tet:THERM_00486810	0.794
Tetrahymena	prot>Q24BW7	GDH_1	tet:THERM_01049200	0.737
Tetrahymena	prot>I7M006	TK_1	tet:THERM_00487030	0.784
Tetrahymena	prot>Q24I41	AP_2	tet:THERM_00569330	0.718
Tetrahymena	prot>I7M8B1	IPPI_2	tet:THERM_00438860	0.767
Tetrahymena	prot>Q23EE3	MDH_1	tet:THERM_01029960	0.713
Paramecium	prot>A0BN51	ADH_1	ptm:GSPATT00030606001	0.772
Paramecium	prot>A0BPK2	ICDH_2	ptm:GSPATT00005218001	0.795
Paramecium	prot>A0CRF7	IMP_D_1	ptm:GSPATT00009689001	0.756
Paramecium	prot>A0BRD6	GDH_1	ptm:GSPATT00031334001	0.718
Paramecium	prot>A0CA73	TK_1	ptm:GSPATT00036470001	0.736
Paramecium	prot>Q7Z111	AP_2	no KEGG	0.772
Paramecium	prot>A0CI91	IPPI_2	ptm:GSPATT00007643001	0.729
Paramecium	prot>A0D8T3	MDH_1	ptm:GSPATT00014396001	0.702
Leishmania	prot>Q4QBD8	ADH_1	lma:LMJF_23_0360	0.755

- **Start > HGT Calculator**

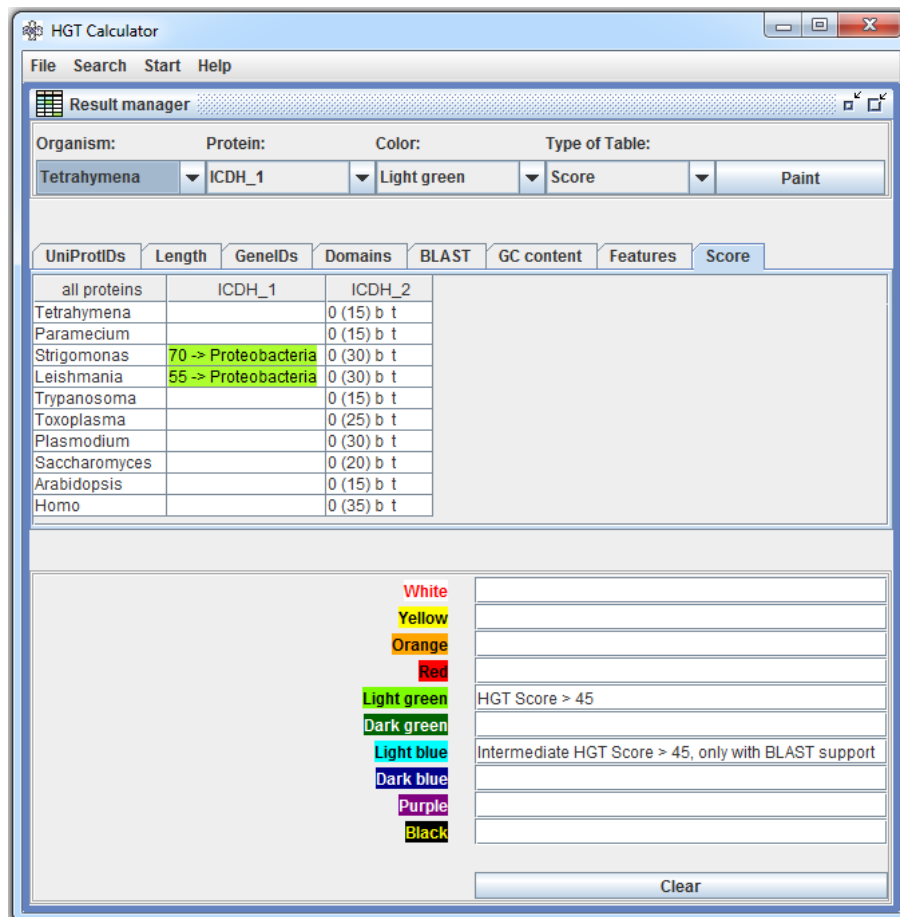
Starts the HGT Calculator

→ Writes the Log file → in workspace\ Data\Download\Log\

→ Writes the Output file → in workspace\ Data\Download\Output\

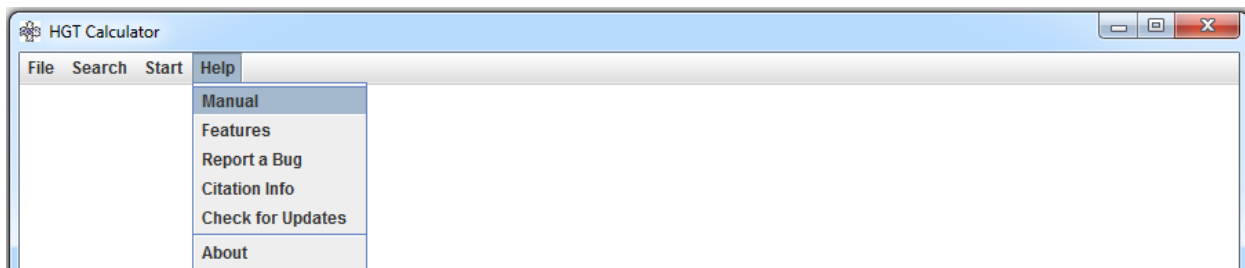
→ Opens Result manager for visualizing the results

Additional parameter: show intermediate score → check box



Scores are set to zero if BLAST results (b) or trees (t) are not convincing. Nevertheless, it is possible that this sequence is a HGT event. Intermediate scores can show this.

If the HGT is an old event which had occurred before different taxonomic lineages separated and two or more species of this group are present in the tree, non of them would be a direct neighbour of a bacterial sequence and therefore without tree support. It would be more likely a subtree of these species clustering together with the bacterial sequence. Another reason could be the occurrence of more than one HGT for this gene. In *Strigomonas culicis*, there are two different homoserine dehydrogenases. One homoserine dehydrogenase sequence seems to be the result of a multiple HGT event from firmicutes into trypanosomatids (low HGT scores), whereas the other one seems to come from an endosymbiont strain of *Strigomonas* (higher HGT score).



- **Help > Manual**
→ Opens this manual.
- **Help > Features**
→ Opens a page about the features with an overview about the different parts of the calculation of the HGT score.
- **Help > Report a bug**
→ Opens an e-mail template to report problems to the author of the application.
- **Help > Citation info**
→ Opens a page with all references on algorithms used by HGT Calculator.
- **Help > About**
→ Opens the about page.
- **File > Quite:**
→ Closes the program.

References for external applications

→ Protein domains

Hunter et al. (2012), InterPro in 2011: new developments in the family and domain prediction database, *Nucleic Acids Research* 40:D306-312.

→ BLAST

Patient et al. (2008), UniProtJAPI: a remote API for accessing UniProt data, *Bioinformatics* 24:1321–1322.

→ CAI calculation

Nakamura Y, Gojobori T, Ikemura T. (2000), Codon usage tabulated from the international DNA sequence databases: status for the year 2000. *Nucleic Acids Research* 28:292.

Xia X. (2007), An Improved Implementation of Codon Adaptation Index. *Evolutionary Bioinformatics* 3:53–58.

→ Trees

Larkin et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947–2948.