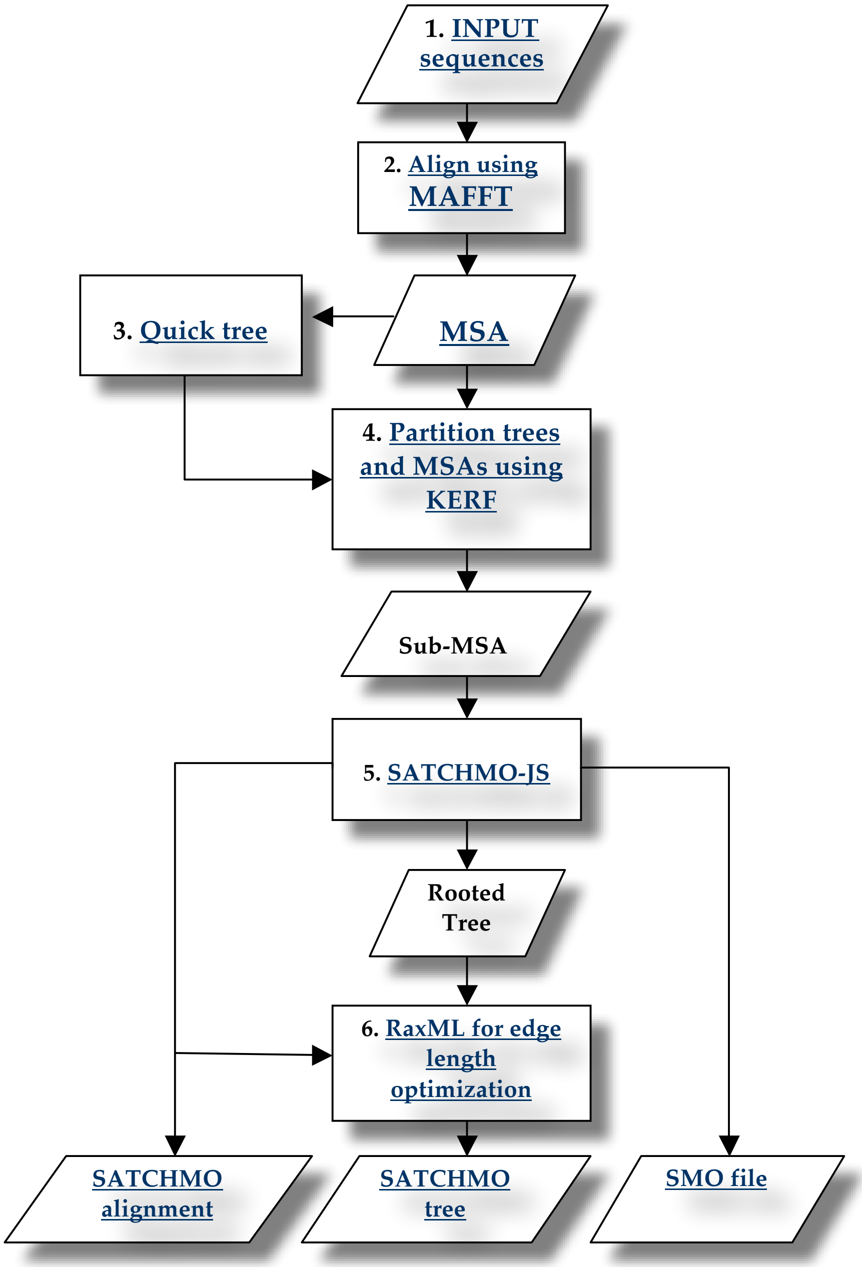
# The SATCHMO Tutorial

SATCHMO is a novel bioinformatics method that simultaneously constructs a phylogenetic tree and multiple sequence alignments (MSA). (*Edgar, R., and Sjölander, K., "SATCHMO: Sequence Alignment and Tree Construction using Hidden Markov models," Bioinformatics. 2003 Jul 22; 19(11):1404-11. Oxford University Press access*). The input is a set of protein sequences in FASTA format. SATCHMO-JS (Jump Start SATCHMO) is a new variant of SATCHMO (see here (http://makana.berkeley.edu/satchmo/help#faq) for an explanation of the differences between SATCHMO and SATCHMO-JS). This Tutorial will help you to understand the SATCHMO server pipeline (http://phylogenomics.berkeley.edu/satchmo-js) , which is one of many tools from Berkeley Phylogenomics Group (http://phylofacts.berkeley.edu/front/).

## The SATCHMO Server Pipeline

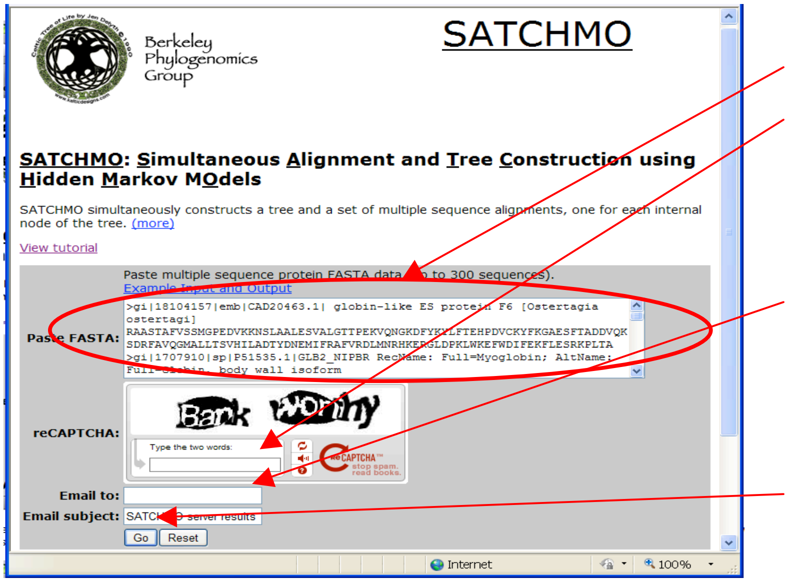
This flow chart describes each step of the SATCHMO server pipeline. You can click on the on the different parts of the pipeline to get a more detailed explanation.



## 1.INPUT Sequences

The SATCHMO server accepts the following as input:

1. Unaligned protein sequences in FASTA format with no special characters allowed (J,O,U, Z).
2. SATCHMO **does not** accept DNA sequences
3. The minimum number of sequences is 4
4. The maximum number of sequences is 300
5. The maximum size of a sequence is 2000 residues



SATCHMO

**SATCHMO: Simultaneous Alignment and Tree Construction using Hidden Markov MOdels**

SATCHMO simultaneously constructs a tree and a set of multiple sequence alignments, one for each internal node of the tree. (more)

View tutorial

Paste multiple sequence protein FASTA data (up to 300 sequences).
Example Input and Output

Paste FASTA:

```
>gi|18104157|emb|CAD20463.1| globin-like ES protein F6 [Ostertagia
ostertagi]
RAASTAFVSSMGPEDVKKNSLAALESVALGTTPEKVQNGKDFYKVLFTEHPDVCKYFKGAESFTADDVQK
SDRFAVQGMALLTSVHILADTYDNEMIFRAFVRDLMNRHKERGLDPKLWKEFWDIFEKFLESRKPLTA
>gi|1707910|sp|P51535.1|GLB2_NIPBR RecName: Full=Myoglobin; AltName:
Full=Globin, body wall isoform
```

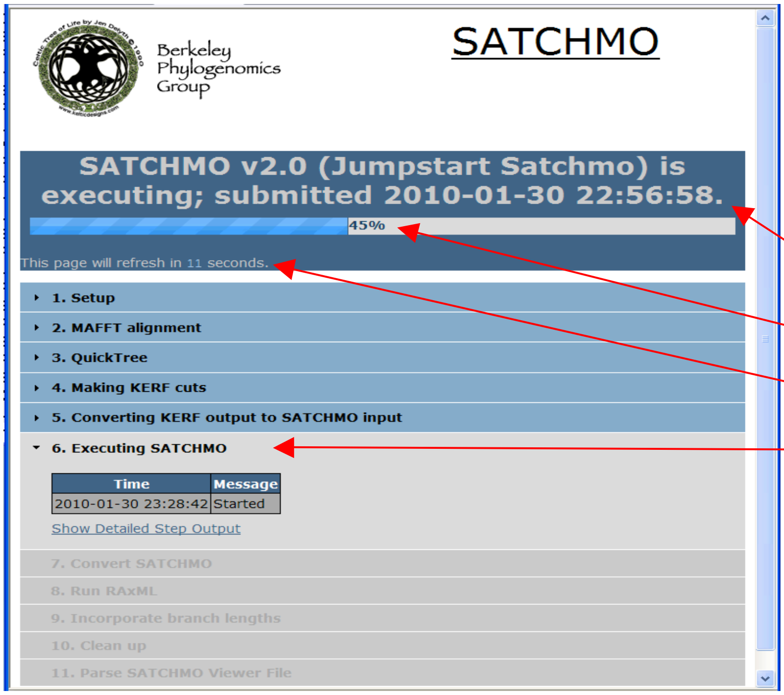
reCAPTCHA: Bank worthy

Type the two words:

Email to:

Email subject: SATCHMO server results

1. Paste your sequence in FASTA format in the text field.
2. Type information in the reCAPTCHA box. This verifies that you are a person and not an automatic script. In the example, the random words are "Bank" and " worthy". These random words change for every new submission.
3. Providing an email address is optional. This way you will get an email with a link to your final results. Or you can bookmark the submitted page and return to view the results later.
4. Click the "Go" button.



SATCHMO

**SATCHMO v2.0 (Jumpstart Satchmo) is executing; submitted 2010-01-30 22:56:58.**

45%

This page will refresh in 11 seconds.

1. Setup
2. MAFFT alignment
3. QuickTree
4. Making KERF cuts
5. Converting KERF output to SATCHMO input
6. Executing SATCHMO

| Time | Message |
|---|---|
| 2010-01-30 23:28:42 | Started |

Show Detailed Step Output

7. Convert SATCHMO
8. Run RAxML
9. Incorporate branch lengths
10. Clean up
11. Parse SATCHMO Viewer File

When the job is submitted successfully you will see this screen that displays the SATCHMO server pipeline. You will be able to track:

1. The date and time of submission
2. The percentage of work completed.
3. How soon before the page is refreshed.
4. Which stage of the pipeline the job is in

## 2.Aligning using MAFFT

MAFFT (Katoh, Misawa, Kuma, Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. 2002 (Nucleic Acids Res. 30:3059-3066).) is an alignment program has produced outstanding results on benchmark datasets.

## MSA (Multiple Sequence Alignment)

The MAFFT algorithm produces a Multiple Sequence Alignment (MSA) in FASTA format that is used as input for two programs:

1. Quick Tree – the alignment is converted to Stockholm format using a Biopython script. Quick tree accepts only Stockholm format alignments.
2. KERF – the MSA is examined to find the minimum pairwise identity between any two sequences in the MSA. The definition of a pairwise identity between two sequences is the ratio of the number of aligned pairs that agree exactly over the number of aligned columns in the MSA. Columns in which both of sequences have a gap should not be counted in either the numerator or the denominator of the ratio. Once the minimum pairwise identity is calculated this value plus 10% is used as input for KERF, to specify the cut-off value.
    - For example, if the MSA has a minimum pairwise identity of 20%, then the KERF cut (next step) will divide the tree into subtrees such that no pair of sequences within any subtree has less than 30% identity.

    If minimum pairwise identity of the MSA created by MAFFT is above 90%, then, the midpoint between the maximum pairwise identity and minimum pairwise identity is used to create the KERF cut.
    - For example, if the MSA has a minimum pairwise identity of 91%, and a maximum pairwise identity of 99%, then the KERF cut will divide the trees into subtrees such that no pair of sequences within any subtree has less than 95% identity.

## 3.Quick Tree

Quick Tree (Kevin Howe, Alex Bateman, and Richard Durbin,,QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics 2002 18: 1546-1547; doi:10.1093/bioinformatics/18.11.1546) is a program that that reconstructs phylogenies using the Neighbor-Joining method. The Quick Tree algorithm takes the previously created Stockholm format MSA and creates a tree in Newick format.

## 4. KERF to build subtrees and MSA partitions

KERF is an in-house method that is not yet published. It takes as input an MSA and a tree and cuts the tree into subtrees. The subtree cuts are based on a user-specified maximum divergence within each subtree that is estimated on the basis of pairwise percent identity in the MSA). The previously generated MAFFT multiple sequence alignment, the minimum Pairwise Identity calculated in the MAFFT step and the Quick Tree generated tree are all used as inputs to KERF. With these inputs, KERF generates cuts of the tree into the fewest possible number of subtrees so that no pair in any subtree has less than the previously stipulated minimum pairwise identity. The output of KERF is a set of sub-MSAs in FASTA format and subtrees. Sometimes many of the sub-MSAs contain columns that are entirely gapped. The solution to this is to mask the sub-MSAs prior to submitting them to SATCHMO. Columns that have an excessive number of gap characters are masked by turning the amino acids in those columns to lower case letters and the corresponding dashes in those positions to dots. SATCHMO should now interpret these positions as generated in an HMM insert state. This will turn any columns, which are at least 70% gaps into lowercase letters and dots. Only the sub-MSAs are used as input for SATCHMO.

## 5.SATCHMO

SATCHMO (Robert C. Edgar and Kimmen Sjolander (2003), SATCHMO: Sequence alignment and tree construction using hidden Markov models, Bioinformatics 19(11), 1404-1411.) simultaneously constructs a tree and a set of multiple sequence alignments, one for each internal node of the tree. Profile hidden Markov models at each node are used to determine, the branching order, the alignment of sequences and the prediction of structurally alignable regions. SATCHMO-JS runs SATCHMO in the jump start mode.

The inputs are:

1. The original submitted sequences
2. The adjusted KERF sub-MSAs.

The outputs are:

1. A SATCHMO alignment - is constructed by aligning the KERF sub-MSAs - Click here to learn how to download, view and interpret the output.
2. A SATCHMO tree – a rooted tree in Newick Format - Click here to learn how to download, view and interpret the output)
3. A SATCHMO SMO file which is the format used in the SATCHMO viewer (only available on the Windows platform) – Click here to learn more about viewing this output.

## 6.RaxML for edge length optimization

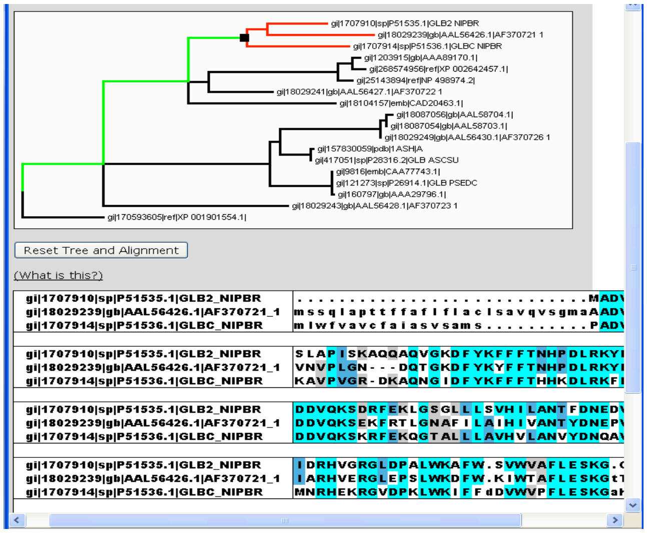
RaxML-VI-HPC (A. Stamatakis, T. Ludwig, and H. Meier. Raxml-iii: A fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics, 21(4):456-463, 2005.) is a program for making phylogenetic trees using the maximum likelihood tree method. The SATCHMO algorithm does not give edge lengths on the rooted tree that it generates. Hence this tree is submitted to the RaxML program along with the SATCHMO alignment to give a tree with edge lengths. The SATCHMO alignment is masked to remove dot characters and lowercase letters prior to being submitted to RaxML.

## Output

There are 4 different ways to view the results - depending upon your needs and personal preference:

1. The Phyloscope Viewer where you can view the generated tree and MSAs already shown on the page.
2. The Jalview Viewer can be used if you have Java on your computer.
3. Downloads allow you to download the alignment and tree files for use in other bioinformatics programs.
4. You can download the SATCHMO combined tree and alignment file for concurrent interactive examination of the MSA and tree. This requires the SATCHMO Viewer (Windows only). SATCHMO viewer software available for download (The author of the free SATCHMO Viewer software is Robert Edgar. We do not maintain this software).

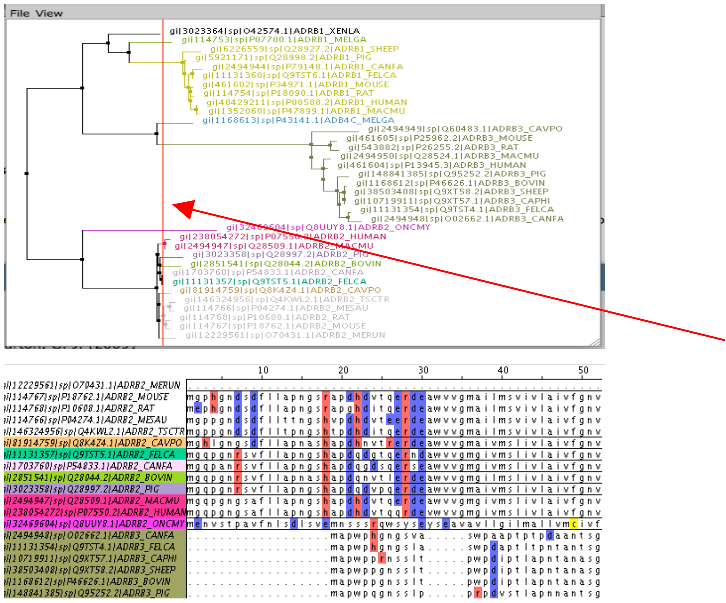
## Phyloscope Viewer


gi|1707910|sp|P51535.1|GLB2 NIPBR
gi|18029239|gb|AAL56426.1|AF370721 1
gi|1707914|sp|P51536.1|GLBC NIPBR
gi|1203915|gb|AAA89170.1|
gi|268574956|ref|XP 002642457.1|
gi|25143894|ref|NP 498974.2|
gi|18029241|gb|AAL56427.1|AF370722 1
gi|18104157|emb|CAD20463.1|
gi|18087056|gb|AAL58704.1|
gi|18087054|gb|AAL58703.1|
gi|18029249|gb|AAL56430.1|AF370726 1
gi|157830059|pdb|1ASH|A
gi|417051|sp|P28316.2|GLB ASCSU
gi|9816|emb|CAA77743.1|
gi|121273|sp|P26914.1|GLB PSEDC
gi|160797|gb|AAA29796.1|
gi|18029243|gb|AAL56428.1|AF370723 1
gi|170593605|ref|XP 001901554.1|
Reset Tree and Alignment
(What is this?)

```
gi|1707910|sp|P51535.1|GLB2_NIPBR          ..........................MAD
gi|18029239|gb|AAL56426.1|AF370721_1       mssqlapttffaflflaclsavqvsgmaAAD
gi|1707914|sp|P51536.1|GLBC_NIPBR          mlwfvavcfaiasvsams..........PAD

gi|1707910|sp|P51535.1|GLB2_NIPBR          SLAPISKAQQAQVGKDFYKFFFTNHPDLRKYI
gi|18029239|gb|AAL56426.1|AF370721_1       VNVPLGN---DQTGKDFYKYFFTNHPDLRKYI
gi|1707914|sp|P51536.1|GLBC_NIPBR          KAVPVGR-DKAQNGIDFYKFFFTHHKDLRKFI

gi|1707910|sp|P51535.1|GLB2_NIPBR          DDVQKSDRFEKLGSGLLLSVHILANTFDNED
gi|18029239|gb|AAL56426.1|AF370721_1       DDVQKSEKFRTLGNAFILAIHIVANTYDNEP
gi|1707914|sp|P51536.1|GLBC_NIPBR          DDVQKSKRFEKQGTALLLAVHVLANVYDNQA

gi|1707910|sp|P51535.1|GLB2_NIPBR          IDRHVGRGLDPALWKAFW.SVWVAFLESKG.
gi|18029239|gb|AAL56426.1|AF370721_1       IARHVERGLEPSLWKDFW.KIWTAFLESKGtT
gi|1707914|sp|P51536.1|GLBC_NIPBR          MNRHEKRGVDPKLWKIFFdDVWVPFLESKGaH
```

The Phyloscope Viewer is the default viewer from the Berkeley Phylogenomics Group Phyloscope software. It gives you a quick preview of the results. If you select portions of the tree (by clicking on an internal node), you will see the MSA below is updated to reflect your selection. Because the SATCHMO algorithm estimates a different MSA for each node in the SATCHMO tree, the MSA for a subtree may have subtle differences from the corresponding rows in the full MSA, which is based on the SATCHMO alignment at the root of the tree. These differences stem from the fact that SATCHMO attempts to identify positions that correspond to "conserved core structure" among the sequences that descend from a node; positions that meet this criterion are displayed in upper case (corresponding to an HMM--Hidden Markov Model--match state) while positions that fail this criterion are displayed in lower case (indicating that they are generated in an HMM insert state). When a dataset includes highly divergent sequences, increasing numbers of characters will display in lower case within alignments toward the root, but the same amino acids may display in upper case at nodes nearer the tips of the tree. In some cases, an entire subfamily (subtree) will contain an extended region found only within that subfamily; these positions might represent a region that was inserted in the most recent common ancestor of the subfamily, so they have a common evolutionary history within the subfamily but not outside of the subfamily.
(HELPFUL HINT: For very large inputs, the Phyloscope program may take some time to process and may give an error).

## Jalview Viewer

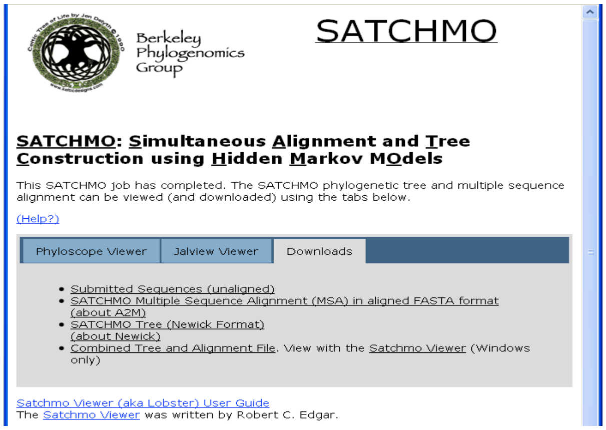

File View
gi|3023364|sp|O42574.1|ADRB1_XENLA
gi|114753|sp|P07700.1|ADRB1_MELGA
gi|6226559|sp|Q28927.2|ADRB1_SHEEP
gi|5921171|sp|Q28998.2|ADRB1_PIG
gi|2494944|sp|P79148.1|ADRB1_CANFA
gi|11131360|sp|Q9TST6.1|ADRB1_FELCA
gi|461602|sp|P34971.1|ADRB1_MOUSE
gi|114754|sp|P18090.1|ADRB1_RAT
gi|48429211|sp|P08588.2|ADRB1_HUMAN
gi|1352060|sp|P47899.1|ADRB1_MACMU
gi|1168613|sp|P43141.1|ADB4C_MELGA
gi|2494949|sp|Q60483.1|ADRB3_CAVPO
gi|461605|sp|P25962.2|ADRB3_MOUSE
gi|543882|sp|P26255.2|ADRB3_RAT
gi|2494950|sp|Q28524.1|ADRB3_MACMU
gi|461604|sp|P13945.3|ADRB3_HUMAN
gi|148841385|sp|Q95252.2|ADRB3_PIG
gi|1168612|sp|P46626.1|ADRB3_BOVIN
gi|38503408|sp|Q9XT58.2|ADRB3_SHEEP
gi|10719911|sp|Q9XT57.1|ADRB3_CAPHI
gi|11131354|sp|Q9TST4.1|ADRB3_FELCA
gi|2494948|sp|O02662.1|ADRB3_CANFA
gi|32469604|sp|Q8UUY8.1|ADRB2_ONCMY
gi|238054272|sp|P07550.2|ADRB2_HUMAN
gi|2494947|sp|Q28509.1|ADRB2_MACMU
gi|3023358|sp|Q28997.2|ADRB2_PIG
gi|2851541|sp|Q28044.2|ADRB2_BOVIN
gi|1703760|sp|P54833.1|ADRB2_CANFA
gi|11131357|sp|Q9TST5.1|ADRB2_FELCA
gi|81914759|sp|Q8K4Z4.1|ADRB2_CAVPO
gi|146324956|sp|Q4KWL2.1|ADRB2_TSCTR
gi|114766|sp|P04274.1|ADRB2_MESAU
gi|114768|sp|P10608.1|ADRB2_RAT
gi|114767|sp|P18762.1|ADRB2_MOUSE
gi|12229561|sp|O70431.1|ADRB2_MERUN

```
                                              10        20        30        40        50
i|12229561|sp|O70431.1|ADRB2_MERUN   ....................................................
i|114767|sp|P18762.1|ADRB2_MOUSE     mgphgndsdfllapngsrapdhdvtqerdeawvvgmailmsvivlaivfgnv
i|114768|sp|P10608.1|ADRB2_RAT       mephgndsdfllapngsrapghditqerdeawvvgmailmsvivlaivfgnv
i|114766|sp|P04274.1|ADRB2_MESAU     mgppgndsdfllttngshvpdhdvteerdeawvvgmailmsvivlaivfgnv
i|146324956|sp|Q4KWL2.1|ADRB2_TSCTR  mgppgndsdflltpngshtpdhditqerdeawvvgmailmsvivlaivfgnv
i|81914759|sp|Q8K4Z4.1|ADRB2_CAVPO   mghlgngsdfllapnashapdhnvtrerdeawvvgmaivmslivlaivfgnv
i|11131357|sp|Q9TST5.1|ADRB2_FELCA   mgqpgnrsvfllapngshapdqdgtqerndawvvgmgivmslivlaivfgnv
i|1703760|sp|P54833.1|ADRB2_CANFA    mgqpanrsvfllapngshapdqgdsqerseawvvgmgivmslivlaivfgnv
i|2851541|sp|Q28044.2|ADRB2_BOVIN    mgqpgnrsvfllapnashapdqnvtlerdeawvvgmgilmslivlaivfgnv
i|3023358|sp|Q28997.2|ADRB2_PIG      mgqpgnrsvfllapngshapdqdvpqerdeawvvgmaivmslivlaivfgnv
i|2494947|sp|Q28509.1|ADRB2_MACMU    mgqpgngsafllapngshapdhdvtqerdeawvvgmgivmslivlaivfgnv
i|238054272|sp|P07550.2|ADRB2_HUMAN  mgqpgngsafllapngshapdhdvtqqrdevwvvgmgivmslivlaivfgnv
i|32469604|sp|Q8UUY8.1|ADRB2_ONCMY   menvstpavfnlsdlsvemnsssrqwsyseyseavavllgilmallvmcivf
i|2494948|sp|O02662.1|ADRB3_CANFA    ..................mapwphgngsva.....swpaaptptpdaantsg
i|11131354|sp|Q9TST4.1|ADRB3_FELCA   ..................mapwphgngsla.....swpdaptltpntantsg
i|10719911|sp|Q9XT57.1|ADRB3_CAPHI   ..................mapwpprnsslt.....pwpdiptlapntanasg
i|38503408|sp|Q9XT58.2|ADRB3_SHEEP   ..................mapwppgnsslt.....pwpdiptlapntanasg
i|1168612|sp|P46626.1|ADRB3_BOVIN    ..................mapwppgnsslt.....pwpdiptlapntanasg
i|148841385|sp|Q95252.2|ADRB3_PIG    ..................mapwpqgnsslp.....prpdvstlapnnantsg
```

The Jalview viewer (www.jalview.org) requires Java. Jalview will display both the MSA and the SATCHMO tree but you cannot examine the changes in alignments at internal nodes of the tree. After you select the Jalview Viewer tab from the results page, click on the "View Multiple Sequence Alignment and Tree" button. Two windows with the tree and MSA are displayed. Some of the features are:

1. The red vertical bar can be moved to cut the tree in different places. The alignment viewer reflects these cuts by highlighting the sequences.

## Downloads



SATCHMO

Berkeley Phylogenomics Group

**SATCHMO: Simultaneous Alignment and Tree Construction using Hidden Markov MOdels**

This SATCHMO job has completed. The SATCHMO phylogenetic tree and multiple sequence alignment can be viewed (and downloaded) using the tabs below.

(Help?)

Phyloscope Viewer | Jalview Viewer | Downloads

- Submitted Sequences (unaligned)
- SATCHMO Multiple Sequence Alignment (MSA) in aligned FASTA format (about A2M)
- SATCHMO Tree (Newick Format) (about Newick)
- Combined Tree and Alignment File. View with the Satchmo Viewer (Windows only)

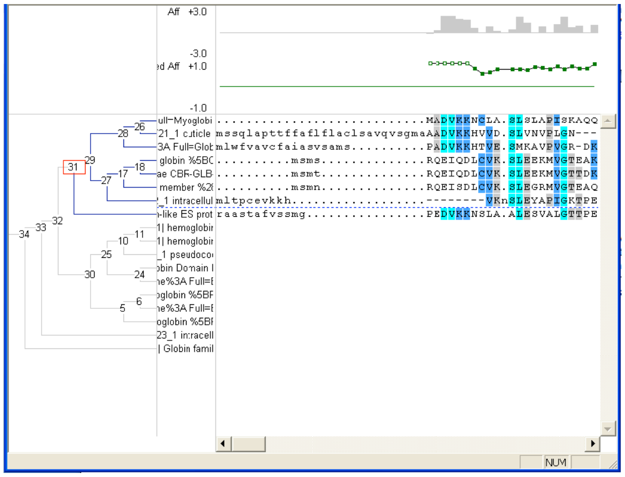
Satchmo Viewer (aka Lobster) User Guide

The Satchmo Viewer was written by Robert C. Edgar.

The Downloads tab of the results page provides the following SATCHMO outputs for download:

- Submitted Sequences (unaligned) - the original set of sequences you submitted
- SATCHMO Multiple Sequence Alignment (MSA) in aligned FASTA format ( aligned FASTA format)
- SATCHMO Tree (Newick Format) - Newick is the standard format for phylogenetic trees that can be used by most phylogenetic tree viewers.
- Combined tree and Alignment File – This is the SMO file which is a file format viewable using SATCHMO viewer.

## SATCHMO Viewer



```
ull-Myoglobi .........................MADVKKNCLA.SLSLAPISKAQQ
21_1 cuticle mssqlapttffaflflaclsavqvsgmaAADVKKHVVD.SLVNVPLGN---
3A Full=Glob mlwfvavcfaiasvsams.........PADVKKHTVE.SMKAVPVGR-DK
 globin %5BC ..........msms.............RQEIQDLCVK.SLEEKMVGTEAK
ae CBF-GLB  ..........msmt.............RQEIQDLCVK.SLEEKMVGTTDK
 member %2(  ..........msmn.............RQEISDLCVK.SLEGRMVGTEAQ
_1 intracellul mltpcevkkh................--------VKnSLEYAPIGKTPE
-like ES prot raastafvssmg...............PEDVKKNSLA.ALESVALGTTPE
```

The SMO file a special file format created by the SATCHMO that can be viewed using the SATCHMO Viewer. The SATCHMO Viewer allows the user to simultaneously view a phylogenetic tree and its alignment. The user can select a node on the tree and instantly view the corresponding sub-alignment. Unfortunately, the SATCHMO Viewer is currently only available for Windows.