

Handbook for AS2TS Homology Modeling Results

(prepared by Korin Wheeler)

Overview: The raw output from AS2TS is designed to provide the maximum amount of information for the user, providing unbiased results for analysis of structure and functional information.

Sample Data Output in the 'Summary File'


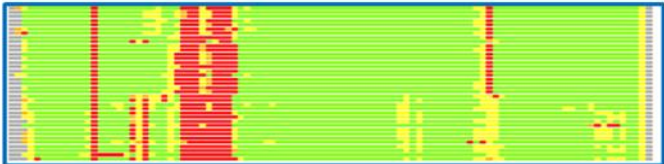
Header info for all proteins modeled

SUMMARY results from the AS2TS modeling

List of submitted protens (FASTA file): [FASTA](#)
 List of created models (AS2TS results): [AS2TS](#)

```

#####
# PROTEIN: Q2219_554_1.5vLI1_02657_1                                104    1 # Category C1
# Length: 104
# Weight: 11888.66
# Similarity: UniProt A3ETKO, BLAST A3ETKO, PDB templates INFO
# Best Pcover Model PDB N AA SISC E-val Seq ID LAL Overlap
# Cat 85.58 M 00 1cm7 A 363 2 1e-24 18.000 89:95 (7-101:132-220)
# Evl 85.58 M 00 1cm7 A 363 2 1e-24 18.000 89:95 (7-101:132-220)
# Sid 84.62 M 02 1a05 A 358 2 1e-21 19.000 88:96 (6-100:126-214)
# Cov 90.38 M 21 1w0d A 337 8 7e-17 19.000 94:100 (2-101:112-205)
# LaL 90.38 M 21 1w0d A 337 8 7e-17 19.000 94:100 (2-101:112-205)
# OvN 90.38 M 21 1w0d A 337 8 7e-17 19.000 94:100 (2-101:112-205)
# OvC 85.58 M 00 1cm7 A 363 2 1e-24 18.000 89:95 (7-101:132-220)
    
```

```

PDB: 1cm7 PDBsum
HEADER OXIDOREDUCTASE 17-MAY-99 1CM7
TITLE 3-ISOPROPYLMALATE DEHYDROGENASE FROM ESCHERICHIA COLI
SOURCE 2 ORGANISM SCIENTIFIC: ESCHERICHIA COLI;
KEYWDS OXIDOREDUCTASE, DEHYDROGENASE, NAD-DEPENDANT ENZYME,
KEYWDS 2 LEUCINE BIOSYNTHETIC PATHWAY
SCOP: c.77 Isocitrate/Isopropylmalate dehydrogenase-like
SCOP: c.77.1 Isocitrate/Isopropylmalate dehydrogenase-like
SCOP: c.77.1.1 Dimeric isocitrate & isopropylmalate dehydrogenases
    
```

Header info for all proteins modeled

The header contains the following important information:

FASTA hyperlinks to a fasta (*.faa) file with sequences for all proteins submitted for modeling.

AS2TS hyperlinks to a list of results for all proteins submitted for modeling (see next page for more info on this file)

The AS2TS file (list of results for all proteins submitted for modeling)

This page contains a short version of the description of a submitted job and modeling results.

The screenshot shows the AS2TS - Model Builder interface. It is divided into three sections indicated by red brackets on the left: Submission information, Hyperlinks, and Modeling info.

Submission information:

- Date: Sun Mar 8 01:40:53 PST 2009
- ADDRESS: adamz@lnl.gov
- AS2TS job number: job_2219_554
- Number of sequences: 560
- Method: FB
- Matrix: BL62
- S-level: 1
- F-split: 700
- N-Libs: 27
- Sequence library: Uniref_90
- Structure library: SEQRES.unique.clean

Hyperlinks:

After job is completed: [All models \(HTML\)](#), [All models \(DIR\)](#), [Summary](#)

Modeling info:

List of modeled proteins:			
Q2219_554_1.SwLII_02657_1	104	1	# Category C1
Q2219_554_2.SwLII_08228_3	139	2	# Category C2
Q2219_554_3.SwLII_08492_1	366	3	# Category C2
Q2219_554_4.SwLII_08970_1	236	4	# Category C1
Q2219_554_5.SwLII_08970_2	238	5	# Category C1
Q2219_554_6.SwLII_08975_2	213	6	# Category C3 No
Q2219_554_7.SwLII_09314_3	100	7	# Category C2
Q2219_554_8.SwLII_09314_4	73	8	# Category C2
Q2219_554_9.SwLII_09490_3	79	9	# Category C1
Q2219_554_10.SwLII_09490_3	79	10	# Category C1
Q2219_554_11.SwLII_09490_5	220	11	# Category C2

Submission Information

- Date date of submission to AS2TS modeling system
- ADDRESS email address of a submission party
- AS2TS job number automatically assigned by the program
- Number of sequences number of sequences submitted
- Method sequence alignment method used for the alignment calculations
- Matrix substitution matrix used by the sequence alignment method
- S-level defines complexity of applied structural modeling procedures:
 - 0 – standard modeling (quick),
 - 1 – using local libraries to enhance homology searches,
 - 2 – enhanced loop building procedures applied (slow)
- F-split if a sequence is over the length specified here, then the sequence will be severed in overlapping fragments.
- N-Libs number of local libraries used for analysis (modeling iterations)
- Sequence library sequence library used for analysis
- Structure library structure library used for analysis

Hyperlinks

Three hyperlinked files are provided in this row:

- All models (HTML) brings you to a list of hyperlinks to the results from all modeling iterations
- All models (DIR) brings you to a directory with all created structural models
- Summary will bring you to the Summary File with the modeling results for each protein (see page 1 and pages 4-8)

Modeling info

The rest of this file is a table with information on each protein entered for modeling. The table is outlined below.

protein modeled protein length system entry number quality category of model

protein modeled	protein length	system entry number	quality category of model
List of modeled proteins:			
Q2219_554_1.5wLII_02657_1		104	1 # Category C1
Q2219_554_2.5wLII_08228_3		139	2 # Category C2
Q2219_554_3.5wLII_08492_1		366	3 # Category C2
Q2219_554_4.5wLII_08970_1		236	4 # Category C1
Q2219_554_5.5wLII_08970_2		238	5 # Category C1
Q2219_554_6.5wLII_08975_2		213	6 # Category C3 NO
Q2219_554_7.5wLII_09314_3		100	7 # Category C2
Q2219_554_8.5wLII_09314_4		73	8 # Category C2
Q2219_554_9.5wLII_09490_3		79	9 # Category C1
Q2219_554_10.5wLII_09490_3		79	10 # Category C1
Q2219_554_11.5wLII_09490_5		220	11 # Category C2

Hyperlinks are provided for –

- Protein model (e.g. “Q2219_554_1.5wLII_02657_1”) This is a hyperlink to the complete results (all models from all iterations) on this protein
- Quality category of model (e.g. “Category_A”) This is a hyperlink to the summary of the modeling results for a given protein – the Summary File.

Sample Data Output for each protein in the 'Summary File'

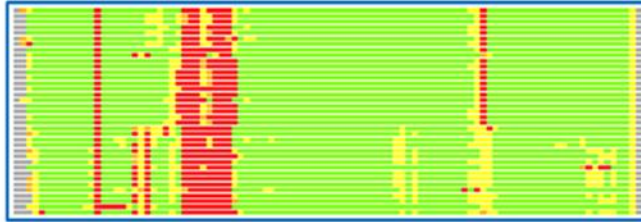
Introductory information

```
#####
# PROTEIN: Q2219 S54 1.5vLII 02657 1 104 1 # Category C1
# Length: 104
# Weight: 11888.66
```

Table on Top 7 models

```
# Similarity: UniProt A3ETK0, BLAST A3ETK0, PDB templates INFO
# Best Pcover Model PDB N_aa SISC E-val Seq_ID LAL Overlap
# Cat 85.58 M 00 1cm7 Å 363 2 1e-24 18.000 89:95 (7-101:132-220)
# Ev1 85.58 M 00 1cm7 Å 363 2 1e-24 18.000 89:95 (7-101:132-220)
# Sid 84.62 M 02 1a05 Å 358 2 1e-21 19.000 88:96 (6-100:126-214)
# Cov 90.38 M 21 1w0d Å 337 8 7e-17 19.000 94:100 (2-101:112-205)
# LaL 90.38 M 21 1w0d Å 337 8 7e-17 19.000 94:100 (2-101:112-205)
# OvN 90.38 M 21 1w0d Å 337 8 7e-17 19.000 94:100 (2-101:112-205)
# OvC 85.58 M 00 1cm7 Å 363 2 1e-24 18.000 89:95 (7-101:132-220)
```

Image of CAT model and Structural comparison of all identified Templates



PDB Information on Structure Templates

```
PDB: 1cm7 PDBsum
HEADER OXIDOREDUCTASE 17-MAY-99 1CM7
TITLE 3-ISOPROPYLMALATE DEHYDROGENASE FROM ESCHERICHIA COLI
SOURCE 2 ORGANISM SCIENTIFIC: ESCHERICHIA COLI;
KEYWDS OXIDOREDUCTASE, DEHYDROGENASE, NAD-DEPENDANT ENZYME,
KEYWDS 2 LEUCINE BIOSYNTHETIC PATHWAY
SCOP: c.77 Isocitrate/Isopropylmalate dehydrogenase-like
SCOP: c.77.1 Isocitrate/Isopropylmalate dehydrogenase-like
SCOP: c.77.1.1 Dimeric isocitrate & isopropylmalate dehydrogenases
```

Introductory Information

```
#####  
# PROTEIN: Q2219 554 1.5wLII 02657 1  
# Length: 104  
# Weight: 11888.66  
# Similarity: UniProt A3ETK0, BLAST A3ETK0, PDB templates INFO
```

Protein name no. amino acids (length) sequence # model quality

Line 1: "PROTEIN:" is followed by the protein name (as titled by user), protein length, the sequence number in the modeling run, and the quality assessment of the protein model generated.

Hyperlinks – *Protein name* hyperlinks to a file with all of the models created for this protein
Category hyperlinks to the Summary File for this protein.

Line 2: "Length:" is followed by the protein length, number of amino acids in protein sequence

Line 3: "Weight:" molecular weight of the protein, as dictated by the amino acid sequence.

Line 4: "Similarity:" each of these three entries are hyperlinks.

- "UniProt ___" clicking here opens a website for the UniProtKB/TrEMBL entry for the protein in question.
- "BLAST ___" clicking here opens the raw BLAST results for the sequence in question.
- "PDB_Templates INFO" clicking here opens a website with information on all PDB templates with similarity to the protein in question.

Table on Top 7 Models

This section is simply a table of results for the models which perform best, as evaluated by 7 different criteria (outlined below):

#	Best	Pcover	Model	PDB	N_AA	SISC	E-val	Seq_ID	LAL	Overlap
#	Cat	85.58	M 00	1cm7 A	363	2	1e-24	18.000	89:95	(7-101:132-220)
#	Evl	85.58	M 00	1cm7 A	363	2	1e-24	18.000	89:95	(7-101:132-220)
#	Sid	84.62	M 02	1a05 A	358	2	1e-21	19.000	88:96	(6-100:126-214)
#	Cov	90.38	M 21	1w0d A	337	8	7e-17	19.000	94:100	(2-101:112-205)
#	LaL	90.38	M 21	1w0d A	337	8	7e-17	19.000	94:100	(2-101:112-205)
#	OvN	90.38	M 21	1w0d A	337	8	7e-17	19.000	94:100	(2-101:112-205)
#	OvC	85.58	M 00	1cm7 A	363	2	1e-24	18.000	89:95	(7-101:132-220)

Line 5: This is the header line for each column in the table, as defined below (from right to left).

- “Best” indicates evaluation criteria.
- “Pcover” is % coverage of the model sequence.
- “Model” clicking on the M hyperlinks to the structural information for the model, clicking on the number begins a download of the PDB file for the created model.
- “PDB” provides hyperlinks for the PDB template used to create the model. The first four letters indicate the PDB ID of the template and the final letter is its chain specification.
- “N_AA” is the number of amino acids in the PDB template sequence.
- “SISC” – number of different sets of coordinates available for a given protein template; provides a hyperlink to additional information about corresponding PDB files.
- “E-val” is the e-value for the alignment of the sequence of interest with the template.
- “Seq_ID” is the % sequence identity of the alignment of the sequence of interest with the template.
- “LAL” is the length of alignment, the number of amino acids extracted from ATOM records from the PDB template : then the length of calculated sequence alignment. Clicking on the numbers will provide additional alignment details (AL2TS format of the alignment).
- “Overlap” provides the exact ranges of amino acid sequences used in sequence alignment, the range of amino acids in the sequence of interest is followed by : the range of aligned amino acids from the template sequence.

Line 6-12: Results for each model that performs best via the criteria below.

- “Cat” is the categorically best model, containing the highest marks in most of the following three criteria: e-value, sequence identity, and sequence coverage.
- “Evl” is the model that has the best e-value.
- “Sid” is the model that has the best sequence identity.
- “Cov” is the model that has the best percent coverage.
- “LaL” is the model that has the best length of alignment (lowest number of gaps).
- “OvN” is the model that has the best coverage/overlap in the N-terminus region.
- “OvC” is the model that has the best coverage/overlap in the C-terminus region.

Image of CAT model



This blue square contains an image of the 'CAT' structural model. The image is colored by secondary structure characteristics; pink indicates an alpha helix and yellow indicates a beta sheet, while blue and white are used in turn and "all others" regions.

NOTE: The modeled portion of the protein may not cover the entire length of the protein sequence. Please check the "Pcover", "LAL", and "Overlap" in the table above to verify the number of amino acids modeled.

Structural comparison of all identified Templates



This plot is a bar representation of regions of structure similarity between the 'CAT' model and other structures from PDB identified by the LGA program. For these comparisons the 'CAT' structural model serves as a frame of reference. Colored bars represent *Calpha - Calpha* distance deviation (from the left (N terminal) to the right (C terminal)) between the model and other PDB structures. Colors represent distances between aligned residues and range from green (below 2Å) to red (above 8Å). This plot gives quick information about other possible PDB structures that can be used as templates for structure modeling, and also highlights regions where they are similar or deviate.

Clicking this image opens a website with detailed information on performed LGA structural comparisons.

PDB Information on Structure Templates

Information on each of the Protein DataBank (PDB) templates used to create the ‘top seven models’ is provided in this section. There may be 0-7 entries in this portion of the results section in the Summary File.

Link to the PDB website link to PDB Sum website date of deposit in PDB PDB id.

```
PDB: 1cm7 PDBsum
HEADER OXIDOREDUCTASE 17-MAY-99 1CH7
TITLE 3-ISOPROPYLMALATE DEHYDROGENASE FROM ESCHERICHIA COLI
SOURCE 2 ORGANISM SCIENTIFIC: ESCHERICHIA COLI;
KEYWDS OXIDOREDUCTASE, DEHYDROGENASE, NAD-DEPENDANT ENZYME,
KEYWDS 2 LEUCINE BIOSYNTHETIC PATHWAY
SCOP: c.77 Isocitrate/Isopropylmalate dehydrogenase-like
SCOP: c.77.1 Isocitrate/Isopropylmalate dehydrogenase-like
SCOP: c.77.1.1 Dimeric isocitrate & isopropylmalate dehydrogenases
```

- **HEADER:** indicates the headline found for this entry in the PDB
- **TITLE:** title of the entry in the PDB
- **SOURCE:** source organism, as deposited in the PDB
- **KEYWDS:** these keywords were identified from the protein description in the PDB and often contain functional information.
- **SCOP:** provides the SCOP ID and fold, superfamily and family information.

NOTE: Not all PDB entries may have assigned IDs from SCOP classification or functional information.