

Documentation for TarO Tutorial, as part of EMBO Practical Course on “Computational Aspects of the Protein Target Selection, Protein Production Management and Structure Analysis Pipeline”

Ian Overton and Geoffrey Barton
School of Life Sciences Research
University of Dundee
Dundee
United Kingdom

Contact: taro@compbio.dundee.ac.uk or geoff@compbio.dundee.ac.uk

Reference: Overton et al. (2008) "TarO: A Target Optimisation System for Structural Biology", *Nucleic Acids Research* 36:W190-W196

Contents

1.0 Introduction	1
1.1 Use Policy	1
1.2 Authors & Citation	1
2.0 TarO User Interface	2
2.1 Home Page	2
2.2 Input Sequences Page	2
2.3 Query Status Table	3
2.4 Annotated Multiple Sequence Alignment	3
2.5 Orthologues Page	4
2.6 Homologues Page	4
2.7 Submit New Query Page	5
2.8 Website Layout	5
3.0 Overview of TarO Processes	6
3.1 TarO Sequence Characterisation Summary	6
3.2 Details of TarO Pipeline	7
3.3 Multiple Sequence Alignment Information	8
3.4 TarO Dataflow	8
4.0 TarO Registration	9
5.0 Questions on the Haemoglobin-beta Query	10
6.0 Hints for Haemoglobin beta query questions	11

1.0 Introduction

TarO analyses a protein sequence by a large number of bioinformatics techniques. These include crystallisation propensity prediction, orthologue searching, and many other sequence-based calculations. Results are tabulated and available via an annotated multiple sequence alignment, that can be edited interactively using Jalview. TarO is focused on Structural Genomics target selection/optimisation, but provides annotations that can be informative to a range of biological questions.

TarO connects to available DAS (distributed annotation system) information via Jalview and links to Dasty2, as well as providing routes to other gateways such as UniProt, COG and the Conserved Domains Database. TarO is hosted by the Barton Group, School of Life Sciences Research, University of Dundee.

1.1 Use Policy

We ask users to wait until the results of their submissions become available before submitting any further sequence queries. Please DO NOT write scripts against the TarO server. If you are thinking about conducting large-scale analyses contact TarO_admin (taro@compbio.dundee.ac.uk). Guest access allows you to try TarO without registering. Although there are no restrictions on guest access, guest results are visible to everyone and will be deleted from the server after a minimum of 8 days.

1.2 Authors & Citation

TarO was developed within the SSPF, primarily by Ian Overton and Geoff Barton with contributions from Jo van Niekerk, Lester Carter, Alice Dawson, David Martin, Scott Cameron, Stephen McMahon, Malcolm White, Bill Hunter and Jim Naismith. Funding was provided by BBSRC under the SPoRT initiative. If you use TarO please cite: Overton et al. (2008) "TarO: A Target Optimisation System for Structural Biology", *Nucleic Acids Research* 36:W190-W196; doi:10.1093/nar/gkn141.

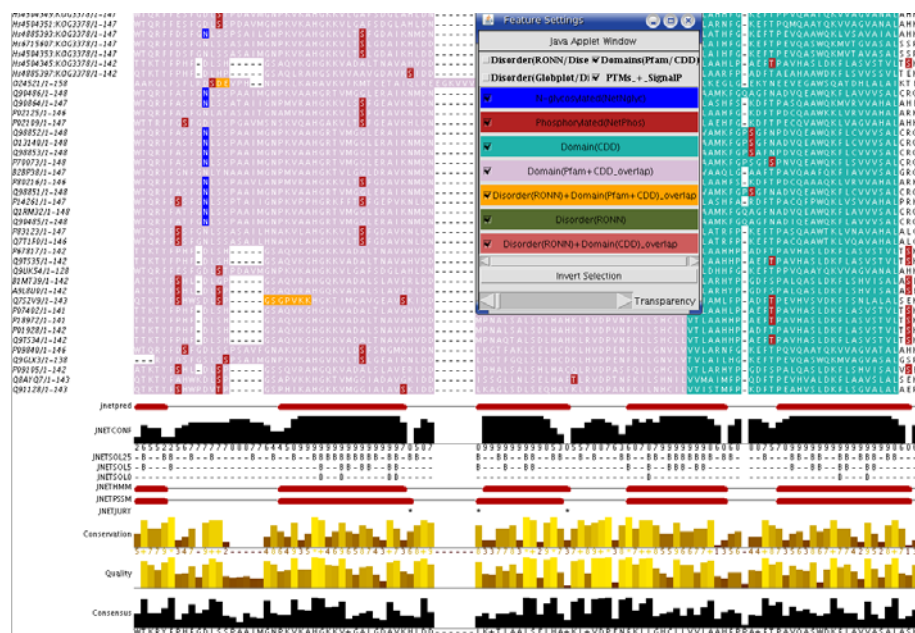
2.3 Query Status Table

Query Status

Step Name	Status
Orthologue searching (BLAST COG)	Results Available
Search Structural Genomics targets (BLAST TargetDB)	Results Available
Protein Disorder Prediction (Globplot & Disembl)	Results Available
RPSBLAST Search CDD, Pfam, COG, KOG & SMART	Results Available
Calculate elementary chemical properties	Results Available
Predict transmembrane regions (TMHMM2)	Results Available
Predict signal peptide (SIGNALP)	Results Available
Multiple sequence alignment (MUSCLE)	Results Available
Search known structures (PSIBLAST PDB)	Results Available
Glycosylation site prediction (NetNglyc, NetOglyc)	Results Available
Phosphorylation site prediction (NetPhos)	Results Available
Protein Disorder Prediction (RONN)	Results Available
Protein Crystallisation Propensity prediction (ParCrys & OB-Score)	Results Available
Secondary Structure Prediction (Jpred)	Results Available
Homologue searching (PSIBLAST UniRef100)	Results Available

On the Input Sequences page there is a table detailing the query progress. The various TarO pipeline stages are summarised in the left column, and the status of each stage is summarised in the right-hand column. The colour of each row reflects the status of each step, according to a ‘traffic lights’ scheme. Orange indicates the step has been started, red indicates the step failed, green indicates the step completes successfully. The table above corresponds to a completed TarO query.

2.4 Annotated Multiple Sequence Alignment



Jalview is used to visualise annotations that can be mapped to residues in the sequence (e.g. phosphorylation sites), other annotations (eg extinction coefficient) are available in the results tables. The Jalview applet provides the facility to start the full Jalview application (on menu click File > View in Full Application). The full Jalview application allows lookup of DAS features and the ability to save alignment files. The multiple sequence alignment (MSA) is constructed using the MUSCLE algorithm, more information on the MSA is given in section 3.3.

SSPF Target Optimisation Utility (TarO) Tutorial

2.5 Orthologues Page



The Barton Group

Home [New Query](#) [Help](#) [Login](#) - contact [TarO_admin](#) (taro@compbio.dundee.ac.uk) to request a private account, free for academic use

Home >> [Input sequences](#) >> [Orthologues](#) >> [Homologues](#)

Query: 657
 Sequence: NTL01AS0002_ACIAD0002-User
 Functional Description: Test_Guest1

Obtain data from this page in [tab-delimited format](#) [html format](#)

Sequence_ID	Links	Organism	ParCrys	ParCrys-Sc	OB	BLASTP Statistics											Sequence statistics								
						eval	%id	Alen	Ost	Qen	Sst	Sen	SeqLen	Mr	GpIclus	pl	GRAVY	SigP	SPconf	#TMH	TMH_span	RO			
NTL01AS0002_ACIAD0002-User	S H C	user	Highly amenable	903e+7	7.08											382	42341	A	5.04	-0.56	0	0.00	0	0-0	0.0
PA0002	S H	<i>Pae</i>	Highly amenable	116e+8	7.78	1e-92	47.26	383	1	382	1	367	367	40695	A	4.89	-0.72	0	0.00	0	0-0	0	0-0	0.0	
RSC3441	S H	<i>Rso</i>	Highly amenable	107e+8	8.25	9e-67	38.28	384	3	382	4	371	371	41058	A	5.72	-0.54	0	0.00	0	0-0	0	0-0	0.0	
NMB1902	S H	<i>Nme</i>	Highly amenable	105e+8	6.25	2e-65	39.01	382	3	382	4	367	367	40854	A	4.85	-0.54	0	0.00	0	0-0	0	0-0	0.0	
NMA0553	S H	<i>NmA</i>	Highly amenable	103e+8	6.25	5e-65	39.01	382	3	382	4	367	367	40930	A	4.78	-0.54	0	0.00	0	0-0	0	0-0	0.0	
dnaN	S H	<i>Eco</i>	Highly amenable	103e+8	4.39	2e-79	44.24	382	1	382	1	366	366	40587	A	5.05	-0.67	0	0.00	0	0-0	0	0-0	0.0	
ECs4636	S H	<i>Ecs</i>	Highly amenable	103e+8	4.39	2e-79	44.24	382	1	382	1	366	366	40587	A	5.05	-0.67	0	0.00	0	0-0	0	0-0	0.0	

This page presents tabulated results for putative orthologues of the input sequence(s), annotated from BLAST searches of the COG database. Results on this page are ordered by predicted crystallisation propensity (ParCrys) and then by BLASTP Expectation value (for the match to the user input sequence). Methodology is described in more detail in section 3.2 below. There are links for each sequence for homologues obtained by a search of UniRef100.

2.6 Homologues Page



The Barton Group

Home [New Query](#) [Help](#) [Login](#) - contact [TarO_admin](#) (taro@compbio.dundee.ac.uk) to request a private account, free for academic use

Home >> [Input sequences](#) >> [Orthologues](#) >> [Homologues](#)

Query: 1025
 Sequence: full-User
 Functional Description: problem sequence

Obtain data from this page in [tab-delimited format](#) [html format](#)

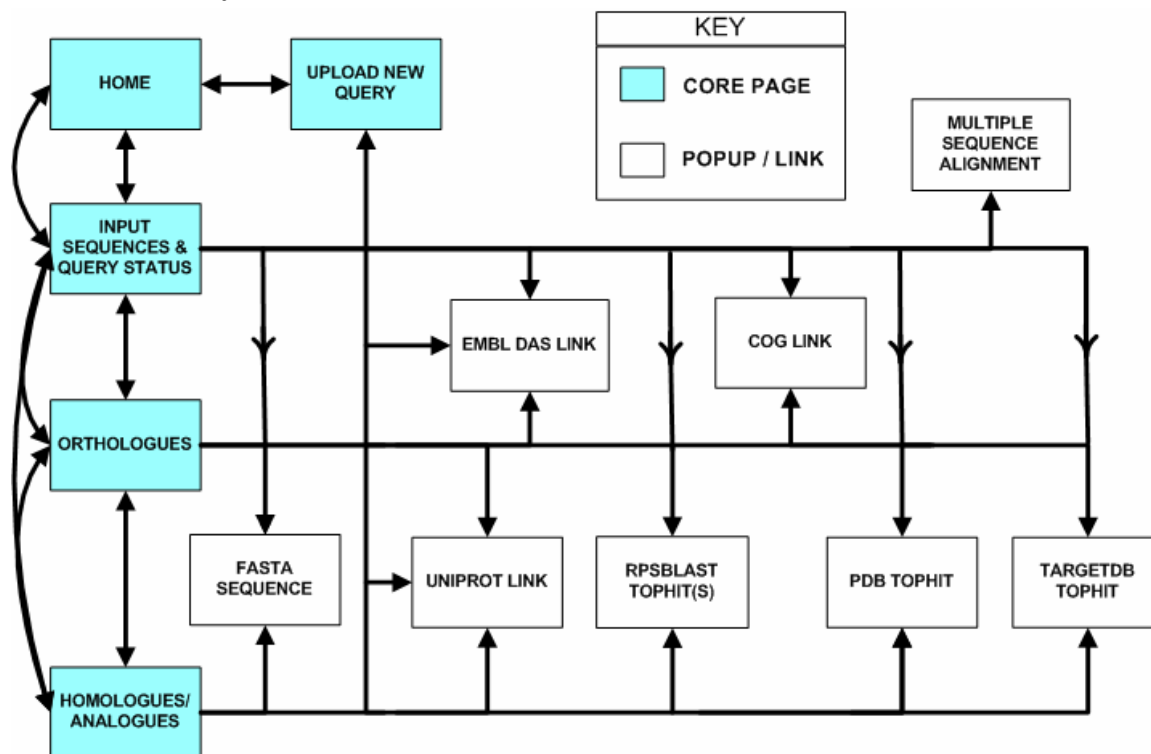
Sequence_ID	Links	Organism	ParCrys	ParCrys-Sc	OB	PSIBLAST Statistics											Sequence statistics								
						eval	%id	Alen	Ost	Qen	Sst	Sen	SeqLen	Mr	GpIclus	pl	GRAVY	SigP	SPconf	#TMH	TMH				
full-User	S	user	Recalcitrant	1.12e+0	-0.27											660	72718	B	8.01	-1.22	0	0.00	0		
UPI0001760581	S D	<i>Danio rerio</i>	High-scoring	9.46e+6	6.13	5e-06	26.19	84	555	636	478	558	1137	127444	A	4.76	-1.04	0	0.00	0					
UPI0000EBC705	S D	<i>Bos taurus</i>	High-scoring	9.00e+6	6.65	1e-06	23.86	88	551	636	761	843	1374	151636	A	6.20	-1.17	0	0.00	0					
UPI0000E21609	S D	<i>Pan troglodytes</i>	High-scoring	8.50e+6	8.65	2e-06	23.86	88	551	636	801	883	1413	157201	A	6.13	-1.12	0	0.00	0					
UPI00016E41C9	S D	<i>Takifugu rubripes</i>	High-scoring	8.44e+6	9.11	2e-07	27.30	84	555	636	487	568	999	113132	A	5.76	-1.15	0	0.00	0					
Q1LR92	S D	<i>Danio rerio</i>	High-scoring	8.39e+6	7.12	5e-07	27.30	84	555	636	477	558	1060	119247	A	5.08	-1.03	0	0.00	0					
UPI0000EB1C3A	S D	<i>Canis lupus familiaris</i>	High-scoring	8.37e+6	6.13	1e-06	23.86	88	551	636	507	590	1122	125942	A	4.80	-1.08	0	0.00	0					
UPI00016E41C9	S D	<i>Takifugu rubripes</i>	High-scoring	8.32e+6	6.56	3e-07	27.30	84	555	636	487	568	1110	125433	A	5.87	-1.19	0	0.00	0					
UPI00006F8AD	S D	<i>Gallus gallus</i>	High-scoring	8.21e+6	6.13	4e-07	26.19	84	555	636	480	561	1095	122905	A	4.78	-1.10	0	0.00	0					
UPI0000A2B94	S D	<i>Canis lupus familiaris</i>	High-scoring	8.10e+6	6.13	1e-06	23.86	88	551	636	478	560	1092	122458	A	4.71	-1.08	0	0.00	0					
		<i>Oryza sativa Japonica</i>																							

This page presents results for putative homologues of the sequence that was clicked on (which could be an input sequence or a COG orthologue). Results on this page are ordered by estimated crystallisation propensity (ParCrys) and then by PSIBLAST Expectation Value. Homologues are gathered using a PSIBLAST search of UniRef100. Methodology is described in more detail in section 3.2 below.

2.7 Submit New Query Page

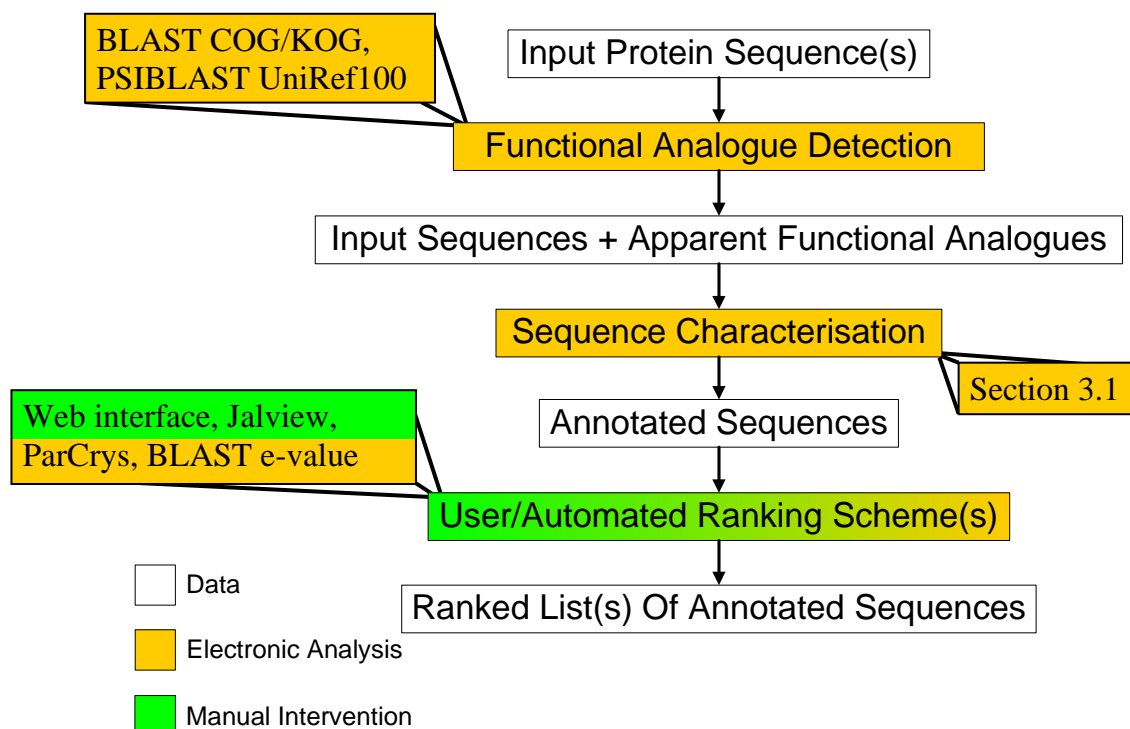
This page is used to start a new TarO query. The query description box allows users to specify a name for the query that is displayed in the home page query summary table. Something that is meaningful to help identify the query to you is therefore recommended. The input is required to be in fasta format and protein sequence. There is the facility to upload an input file, or to paste a fasta format sequence into the large box. This page also has a field to specify the maximum number of sequences to include in the Multiple Sequence Alignment, the default is 100. If too many sequences are included, the alignment may become rather “gappy”.

2.8 Website Layout



The above figure provides a summary of the TarO website layout, to help with finding your way around

3.0 Overview of TarO Processes



3.1 TarO Sequence Characterisation Summary

Analysis Protocol	Brief description
BLASTP TargetDB	Search Structural Genomics targets
PSIBLAST/BLASTP PDB	Search known structures
RPSBLAST CDD, Pfam, COG/KOG, SMART	Search domain profiles
MUSCLE	Multiple sequence alignment
Disembl, RONN, GlobPlot	Protein disorder/order prediction
SignalP	Signal peptide prediction
TMHMM2	Transmembrane region prediction
NetNGlyc, NetOGlyc	Glycosylation site prediction
NetPhos	S/T/Y phosphorylation sites (eukaryotic)
JNET	Secondary structure prediction
Calculate elementary sequence statistics	Sequence length, #M/C/H, Isoelectric point, Molecular weight, hydrophobicity, Extinction coefficient (280nm)
OB-Score, ParCrys, Gravy-pI clustering	Estimate crystallisation propensity

3.2 Details of TarO Pipeline

1 User input sequence(s) searched against the COG database using BLASTP (thresholds coupling sequence identity with alignment length as defined in Rost (1999) Protein Eng. 12:85-94). The topscoring matched COG sequence is used to assign a COG cluster to the input sequence and COG sequences from that COG cluster (of putative orthologues) are thus associated with the user input sequence. Sequences within an assigned COG cluster are displayed if the BLASTP evalue is $1e-3$ or better.

2 All user and associated COG sequences are searched against UniRef100 using PSIBLAST (3 iterations, thresholds: alignment length 30 residues and evalue better than $1E-3$) The resultant matches are assigned to the relevant query (ie user or COG) sequence. The topscoring match from the first iteration is designated the "Uniref Top hit" for each sequence, thus this is the equivalent of a BLASTP search.

3 The User, COG and UniRef100 sequences are analysed in several steps, as follows:

- a) PSIBLAST PDB database (to identify matches to known molecular structure via 'Rost' thresholds)
- b) BLAST TargetDB database (to identify matches to Structural Genomics targets)
- c) RPSBLAST Search domain databases (these are CDD, Pfam, SMART, COG and KOG)
- d) Calculate chemical properties (i.e. pI, Mr, GRAVY, #His, #Met, #Cys, sequence length, extinction coefficient)
- e) SignalP prediction of signal peptide (only first 70 a.a. are examined and results are filtered by the criteria HMM probability threshold ≥ 0.7)
- f) Multiple alignment (MUSCLE). This includes up to 100 sequences by default, which are selected as follows:

Sequences are ordered in the following sections:

- i) user sequence(s) (displayed at the top of the MSA).
- ii) COG sequences
- iii) UniRef100 Sequences

Within each of sections ii) and iii), sequences are displayed in order of similarity to the user sequence(s) (as estimated by (PSI)BLAST expectation values).

- g) Protein disorder/order prediction (using RONN, Globplot and Disembl)
- h) Glycosylation site prediction (O and N-linked, using NetOglyc and NetNglyc)
- i) Phosphorylation site prediction (NetPhos)
- j) Transmembrane region prediction (TMHMM2)
- k) Crystallisation propensity prediction (ParCrys and OB-Score)
- l) Secondary structure prediction (JNET)

3.3 Multiple Sequence Alignment (MSA) Information

Clicking on the button to "View Multiple Sequence Alignment Annotated with...." starts the Jalview applet, displaying a window with the MSA, and a window entitled "Feature Settings". The full Jalview application can be started from within the applet for additional functionality (click 'File'->'View in Full Application'). The MSA was constructed with the MUSCLE algorithm, including sequences that have a BLAST match to the query sequence of $1E-20$ or better. Sequences are excluded if their sequence length is more than 125% of the query sequence length.

The "Feature settings" window can be used to select which annotations to display and the order of precedence for displaying the annotations. The tick-boxes toggle the display of the groups (in the area at near the top of the window) and features (in the 'scroll-able' part of the window). The features are presented in coloured bars that indicate the colour displayed on the MSA when representing annotation of the specified feature. Unselecting a group and then reselecting it will move it to the top of the display order (ie on top of any other selected groups). There are currently 5 groups of annotation on the MSA, however simultaneous

SSPF Target Optimisation Utility (TarO) Tutorial

display of all groups can be confusing! Therefore we strongly suggest that you customise the display of groups using the "Feature Settings" window (described above). The groups are:

i) PTMs_+_SignalP (PostTranslational Modifications and Signal Peptide)

PTMs include:

N & O glycosylation (predicted by NetNglyc and NetOglyc)

Phosphorylation (predicted by NetPhos)

Signal Peptide (predicted by SignalP) is also included here.

NOTE if a signal peptide is not predicted, any predicted glycosylation sites are likely to be wrong!

ii) Domains(Pfam+CDD) & Disorder (RONN)

Domain annotations are from RPSBLAST searching Pfam and CDD profiles.

Disorder is predicted by RONN

iii) TM_regions

TransMembrane regions are predicted by TMHMM2

iv) Disorder (RONN+Disembl)

This group combines protein disorder predicted by RONN and Disembl.

The Disembl "HotLoops" and "REM465" predictions are displayed, however the "COILS" predictions are not displayed.

v) Disorder (Globplot+Disembl)

This group combines protein disorder predicted by Globplot and Disembl.

The Disembl "HotLoops" and "REM465" predictions are displayed, however the "COILS" predictions are not displayed.

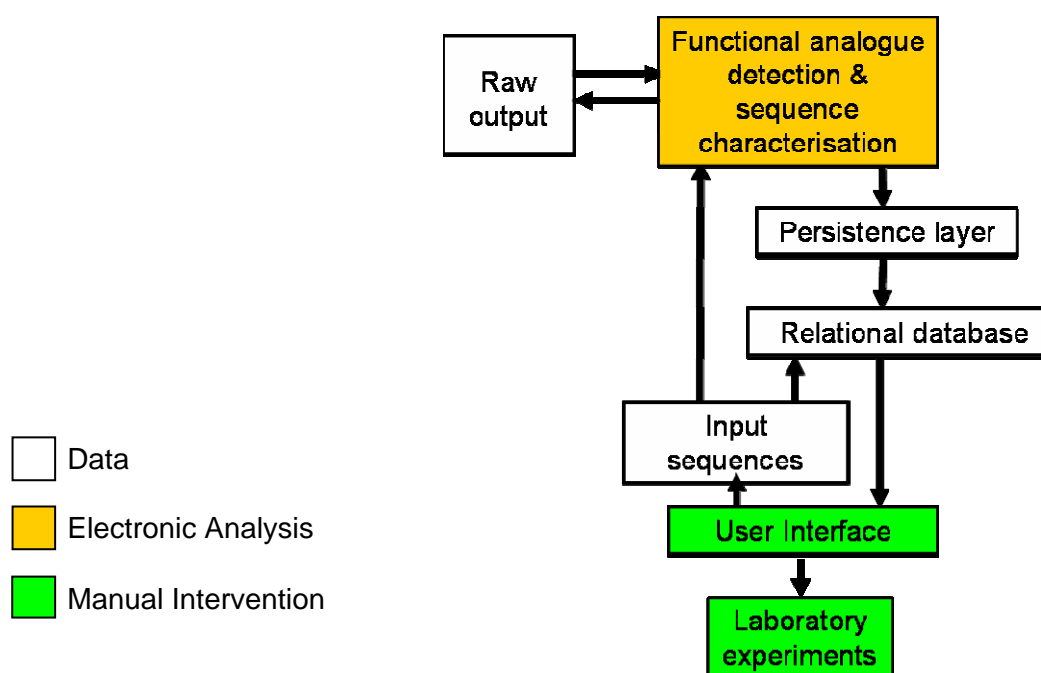
We suggest the following combinations of displayed groups (with i (and iii) displayed on top of the other groups):

i, ii & iii (This is the default, though the group will not be displayed if there is no annotation in that group. Also, the group display order may need tweaking (eg to bring PTMs to the top)).

i, iii & iv

i, iii & v

3.4 TarO Dataflow



4.0 TarO Registration

In order to request a private TarO account contact TarO admin. (taro@compbio.dundee.ac.uk) giving the following details:

- 1) Name
- 2) Institution
- 3) Department
- 4) Country
- 5) Contact email
- 6) Position (e.g. postdoc)
- 7) Supervisor/group name (if applicable)

Please note: Private accounts are currently only available to academic researchers and students working in non-profit institutions such as universities and research institutes.

5.0 Questions on the Haemoglobin-beta Query

1. What are the crystallisation propensity predictions for:
 - (a) the Input sequence
 - (b) the top-ranked KOG sequence (on the 'Orthologues' page)?
2. What are the organisms of the 3 top-ranked homologues from searching the input sequence against UniRef100?
3. What is the Jpred predicted secondary structure content of the input sequence?
4. What is the identifier code for the input sequence's top PDB hit?
5. What region of the input sequence matches to Pfam?
6. Which KOG sequence (on the 'Orthologues' page) has the greatest number of NetPhos predicted phosphorylation sites?
7. What are the crystallisation propensity predictions for the top-scoring homologue of the sequence from question 6?
8. Using the DASTY link, what literature reference is available for the sequence from question 7?
9. What residues are predicted to be phosphorylated for the top-ranked KOG sequence (on the 'Orthologues' page)
10. Which of the sites from question 9 agree with annotations given by the UniProt and Dasty links?
11. By inspecting the annotated multiple alignment, accessible from the 'Input Sequences' page, what residue of the input sequence is the first to be predicted in helical conformation by Jpred?

Hints are given on the next page if you need some help finding the answers

6.0 Hints for Haemoglobin beta query questions

Note that all headings in the results table link to the help page. Also, you may sometimes prefer to open links as a new window. Hints for the tutorial questions are given below.

1. TarO currently gives results for 3 crystallisation propensity prediction methods: ParCrys, the OB-Score and Gravy-pI
2. To navigate to UniRef100 homologues click on the 'H' link near to the input sequence identifier on the 'Input Sequences' page
3. Jpred secondary structure predictions are given in the table on the 'Input Sequences' page under columns 'Jpred_H' (helix) and 'Jpred_E' (extended)
4. The PDB code is available by scrolling to the 'PDB Top Hit' section of the table on the 'Input Sequences' page and clicking the 'P' link
5. Click on the 'RPSBlast results' link in the table on the 'Input Sequences' page, Alternatively, this information may also be obtained from the TarO-annotated multiple sequence alignment, viewed from Jalview.
6. Navigate to the KOG orthologues from the 'O' link near to the input sequence identifier on the 'Input Sequences' page
7. Navigate to the UniRef100 homologues from the 'H' link near to the relevant sequence identifier
8. The link to Dasty is in the 'UniRef Top Hit' section of the table (click on the 'D')
9. More details on phosphorylation site predictions can be obtained from the 'Sequence statistics' section of the table, via the link under the 'NetPhos' column
10. The link to UniProt is in the 'UniRef Top Hit' section of the table (click on the 'U'), the link to Dasty is in the 'UniRef Top Hit' section of the table (click on the 'D')
11. The Jpred predictions appear at the bottom of the alignment. A column can be selected in Jalview by left-clicking the ruler directly above the alignment: this may help you identify the exact position of the helix end in the input sequence.