Deliverable D7.3

Project Title:	World-wide E-infrastructure for structural biology	
Project Acronym:	West-Life	
Grant agreement no.:	**675858**	
Deliverable title:	ProteinCCD with construct scoring and ranking	
WP No.	7	
Lead Beneficiary:	2: NKI AVL	
WP Title	Joint research	
Contractual delivery date:	Month 24	
Actual delivery date:	Month 24	
WP leader:	A. PERRAKIS	NKI AVL
Contributing partners:	N/A	

Deliverable written by Anastassis Perrakis

Contents

1 Executive summary

The Protein Crystallographic Construct Design [1] ProteinCCD (https://xtal.nki.nl/ccd/) software we previously described in deliverable 7.2 aims to increase the efficiency of researchers producing soluble protein in amounts suitable for structural studies, by facilitating the design of several truncation constructs of a protein under investigation. ProteinCCD functions as a meta server that collects information from (web-based) external software that predicts from sequence secondary structure, disorder, coiled coils, transmembrane segments, domains and domain linkers. Viewing the protein sequence annotated with the prediction results allows users to interactively choose possible starts and ends for suitable protein constructs and designing primers needed for PCR amplification. ProteinCCD outputs a comprehensive view of all constructs and all primers needed for bookkeeping and/or ordering of the designed primers.

The functionality of ProteinCCD has been extended under 7.1 to a new computational platform allowing a more interactive and efficient interface to the user, and providing new analysis options. These include parallel processing of server requests, more efficient interface for construct design, more cloning methods, an extended collection of existing vectors, local execution of some algorithms for improving response time, new servers for meta-analysis, easy bookkeeping, and better data security.

Working towards the goal of this deliverable, to provide construct scoring and ranking we implemented several features to reach this goal. Automated alignments of the “work” protein to orthologues present in typical model species, are now provided to the user to facilitate better choices for constructs. All constructs can now be given to the users not only as the “native” protein sequence (as before) but also in the specific context of the cloning vector used for production, including purification tags, and the sequence after enzymatic cleavage of the tags. This is important, as each proteins version has different properties. The molecular weight, isoelectric point, and absorption coefficient for every construct is also computed, enabling the users to understand the properties of the produced proteins.

The final goal to rank the chances of successfully producing the proteins, is realized by assessing the chances to produce soluble proteins for each protein. A score from 0-1 is given by different servers, and provided to the users. The constructs can then be ranked according to these scores.

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	**Provide analysis solutions for the different Structural Biology approaches**	X	
2	**Provide automated pipelines to handle multi-technique datasets in an integrative manner**		X
3	**Provide integrated data management for single and multi-technique projects, based on existing e-infrastructure**	X	
4	**Foster best practices, collaboration and training of end users**	X	

3 Detailed report on the deliverable

3.1 Background

The production of soluble proteins in amounts suitable for structural studies has been a common bottleneck in structural biology and structural genomics alike. Cloning techniques are high-throughput, inexpensive and compatible with robotic implementations, allowing parallel construction of tens of expression constructs for each protein under study: that is a standard practice in many labs. Expression constructs can be designed based on experimental information, but also by computationally. Despite significant progress in sequence analysis, there is no definitive method of choice and submitting many queries to different servers, and subsequent collection of the results of different methods, is the normal laboratory practice. The researcher then typically decides what are promising domain boundaries for the protein in hand, and designs oligonucleotides to be used for PCR-based amplification of all these fragments. At this stage a trivial but time consuming additional bottleneck is encountered: the protein-based analysis has to be transformed back to the DNA sequence. Although the task is by all means trivial, it is time consuming and error prone, since the direct mapping between protein and DNA sequence is lost in the analysis step.

The ProteinCCD (https://xtal.nki.nl/ccd/), introduced for the first time in 2008 (Mooij et al., 2009), is a meta-server to cater for the needs of the above tasks. It automates sequence analysis, and provides interaction with the user for the optimal design of protein constructs that are good candidates for structural analysis. The collection of sequence analysis tools, includes servers for the prediction of secondary structure, disorder, coiled coils, transmembrane segments, domains and domain linkers. A clear and concise view of the protein sequence annotated with the prediction results allows users to interactively choose possible starts and ends for suitable protein constructs. ProteinCCD can help designing the primers needed for PCR amplification of all constructs, as the required user input is the DNA and not the protein sequence. ProteinCCD outputs a comprehensive view of all constructs and all primers needed for bookkeeping and/or ordering of the designed primers.

The application was originally delivered as a Java applet to the client. As the Java approach however, had security issues and Java applets cannot be run on all devices, we previously re-implemented ProteinCCD as a web application, within the West-life project. We used a Python Flask backend and Biopython for the backend. The frontend relies on Bootstrap and Javascript. The new implementation eliminated security concerns, making the application available to any device able to run a web browser, and can be extended with new functionality easily. The user interface has been given a modern and more functional look, emphasising on the most commonly used features based on user feedback. Elimination of all Java dependencies eliminates loading times and needs for certificates validation, improving significantly the user experience.

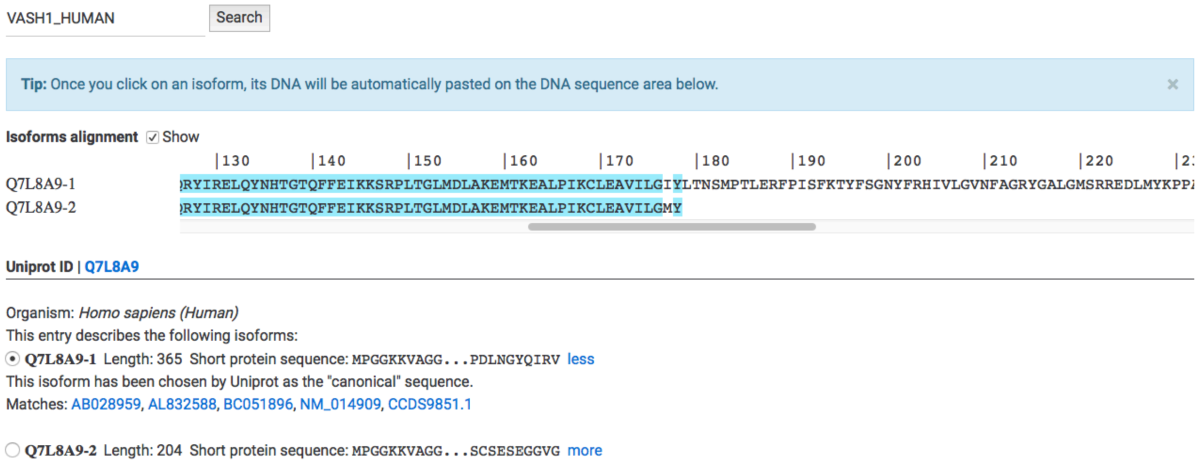
The new design had allowed several analysis options to be implemented. These included parallel processing of server requests, an improved primer design interface implementing the NKI Ligation Independent Cloning (LIC) system (Luna-Vargas et al., 2011), restriction based cloning, and other custom laboratory collections (Celie et al., 2016). In addition the internal database collection, is checked on the fly for vectors with compatible ligation sequences, a small user database of most common restriction enzymes is provided, local execution of some algorithms improved response time, new servers for the meta-analysis were added, and saving of predictions, primers and resulting peptides has been implemented. These changes set the ground for delivering the work reported here.

Figure 1. The entry page of the new CCD server

3.2 Current work

3.2.1 Fetching DNA by UNIPROT accession number and choosing isoforms

The user can now directly enter a UNIPROT accession code for his molecule of interest (Figure 1). ProteinCCD will search for this entry, and if it exists it will present the user with a list of all the available isoforms that are annotated. The user can choose the appropriate isoform, and then corresponding DNA sequence will be automatically used for subsequent analysis. The user can select an alignment between isoforms and request more detailed information for any given isoform.



Tip: Once you click on an isoform, its DNA will be automatically pasted on the DNA sequence area below.

Isoforms alignment ☑ Show

```
           |130      |140      |150      |160      |170      |180      |190      |200      |210      |220      |2
Q7L8A9-1  RYIRELQYNHTGTQFFEIKKSRPLTGLMDLAKEMTKEALPIKCLEAVILGIYLTNSMPTLERFPISFKTYFSGNYFRHIVLGVNFAGRYGALGMSRREDLMYKPP
Q7L8A9-2  RYIRELQYNHTGTQFFEIKKSRPLTGLMDLAKEMTKEALPIKCLEAVILGMY
```

Uniprot ID | Q7L8A9

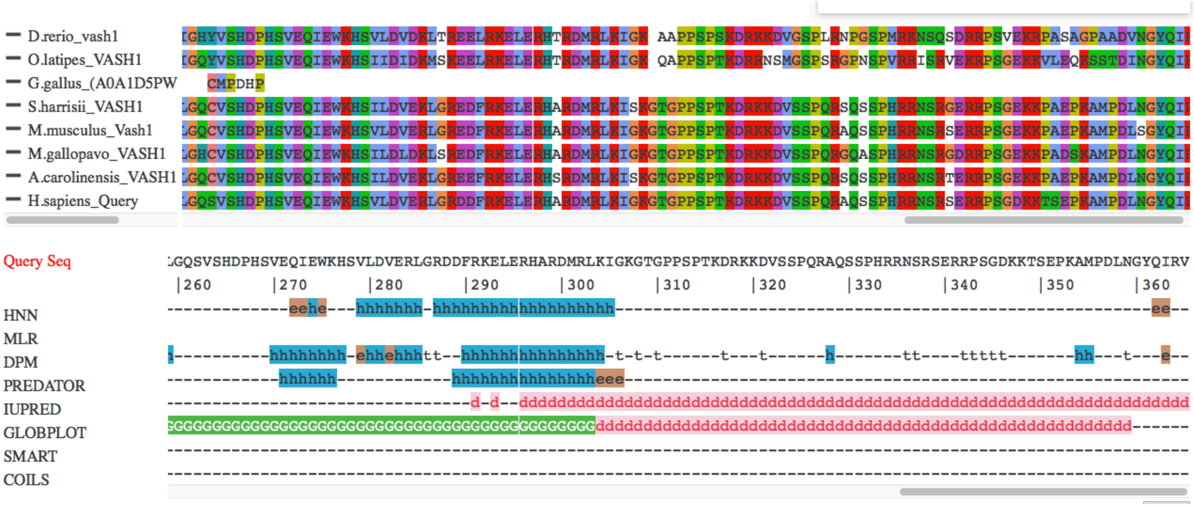
Organism: *Homo sapiens (Human)*
This entry describes the following isoforms:
● **Q7L8A9-1** Length: 365 Short protein sequence: MPGGKKVAGG...PDLNGYQIRV less
This isoform has been chosen by Uniprot as the "canonical" sequence.
Matches: AB028959, AL832588, BC051896, NM_014909, CCDS9851.1

○ **Q7L8A9-2** Length: 204 Short protein sequence: MPGGKKVAGG...SCSESEGGVG more

Figure 2. Choice of isoforms from UNIPROT entries

3.2.2 Alignment of homologues facilitates better user choices

Previous software versions were allowing the choices to be made based on secondary structure predictions and other servers that characterise the sequence. In this version the user is presented with a multiple sequence alignment, that allows to make better-informed choices. An example is depicted in Figure 2. Here, the prediction for the C-terminus suggests that the protein is disordered, and a good construct choice would be to ignore this sequence for expression experiments. However, in the alignment it can be immediately seen that this disordered region is well-conserved, suggesting an important functional role. This observation can influence experimental design but can also allow generating a functional hypothesis for the role of this region, in e.g. participating in crucial protein-protein interactions.



```
D.rerio_vash1          IGHYVSHDPHSVEQIEWKHSVLDVDKLTREELRKELERHTRDMRLKIGK AAPPSPSKDRKKDVGSPLRNPGSPMRKNSQSDRRPSVEKRPASAGPAADVNGYQI
O.latipes_VASH1        IGQYVSHDPHSVEQIEWKHSIIDIDKMSKEELRKELERHTRDMRLKIGK QAPPSPTKDRRNSMGSPSRGPNSPVRRISRVEKRPSGEKKVLEQKSSTDINGYQI
G.gallus_(A0A1D5PW        CMPDHP
S.harrisii_VASH1       LGQCVSHDPHSVEQIEWKHSILDVEKLGREDFRKELERHARDMRLKISKGTGPPSPTKDRKKDVSSPQRSQSSPHRRNSRGERRPSGEKKPAEPKAMPDLNGYQI
M.musculus_Vash1       LGQCVSHDPHSVEQIEWKHSVLDVERLGREDFRKELERHARDMRLKIGKGTGPPSPTKDRKKDVSSPQRAQSSPHRRNSRSERRPSGEKKPAEPKAMPDLSGYQI
M.gallopavo_VASH1      LGHCVSHDPHSVEQIEWKHSILDLDKLSREDFRKELERHARDMRLKIGKGTGPPSPTKDRKKDVSSPQRGQASPHRRNSRGDRRPSGEKKPADSKAMPDLNGYQI
A.carolinensis_VASH1   LGQCVSHDPHSVEQIEWKHSILDVEKLGREEFRKELERHSRDMRLKISKGTGPPSPTKDRKKDVSSPQRSQSSPHRRNSRTERRPSGEKKPAEPKAMPDLNGYQI
H.sapiens_Query        LGQSVSHDPHSVEQIEWKHSVLDVERLGRDDFRKELERHARDMRLKIGKGTGPPSPTKDRKKDVSSPQRAQSSPHRRNSRSERRPSGDKKTSEPKAMPDLNGYQI

Query Seq    LGQSVSHDPHSVEQIEWKHSVLDVERLGRDDFRKELERHARDMRLKIGKGTGPPSPTKDRKKDVSSPQRAQSSPHRRNSRSERRPSGDKKTSEPKAMPDLNGYQIRV
              |260      |270      |280      |290      |300      |310      |320      |330      |340      |350      |360
HNN          -------------eehe---hhhhhhh-hhhhhhhhhhhhhhhhhhh--------------------------------------------------------ee--
MLR
DPM          h----------hhhhhhhh-ehhehhhtt--hhhhhhhhhhhhhhh-t-t-t------t---t------h-------tt----ttttt-------hh---t---e--
PREDATOR     ------------hhhhhh------------hhhhhhhhhhhhhhheee-----------------------------------------------------------
IUPRED       --------------------------------d-d--dddddddddddddddddddddddddddddddddddddddddddddddddddddddddddddddddddddd
GLOBPLOT     GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGdddddddddddddddddddddddddddddddddddddddddddddddddddddddd------
SMART        -----------------------------------------------------------------------------------------------------------
COILS        -----------------------------------------------------------------------------------------------------------
```

Figure 3. The multiple alignment on the top shows excellent conservation of the C-terminus, while the bottom part show that the structure is most likely disordered.

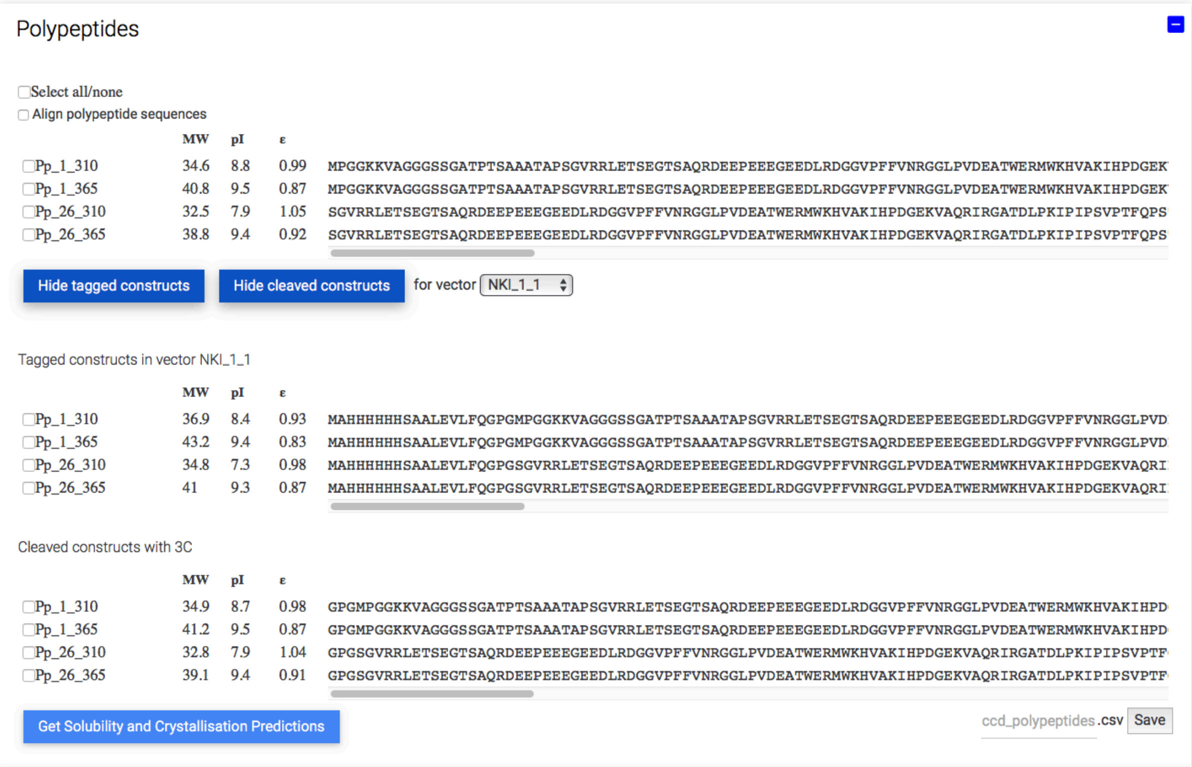
3.2.2 Detailed definition of produced proteins

ProteinCDD outputs the polypeptide sequence of the specific constructs. Previously, this was just the chosen protein sequence. However, in heterologous recombinant expression experiments, in most cases additional “tag” sequences are included fused to the protein, and these can be cleaved with specific proteases to produce “un-tagged” proteins, often leaving “residual” sequences to the expressed protein. The new version of ProteinCCD gives the user the option to display the native protein sequence, the tagged constructs in the context of the chosen vector, and the constructs following tag cleavage. This is important information to document future experimental design.

3.2.3 Physicochemical properties of produced proteins

Experimentalists always need to know basic details for the produced proteins: typically, the molecular weight, isoelectric point, and absorption coefficient (ε). These would have to be calculated separately, cutting and pasting all protein versions. Proteins CCD, now automatically calculates all these parameters for all constructs (native, tagged, and cleaved). It is important to note, that often users presume that the e.g. the native and tagged construct would have very similar parameters and use e.g. the same value for ε; as demonstrated in Figure 3 this assumption can be wrong, especially for His-tagged constructs.

-



Polypeptides

□ Select all/none

□ Align polypeptide sequences

	MW	pI	ε	
□ Pp_1_310	34.6	8.8	0.99	MPGGKKVAGGGSSGATPTSAAATAPSGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVDEATWERMWKHVAKIHPDGEK
□ Pp_1_365	40.8	9.5	0.87	MPGGKKVAGGGSSGATPTSAAATAPSGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVDEATWERMWKHVAKIHPDGEK
□ Pp_26_310	32.5	7.9	1.05	SGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVDEATWERMWKHVAKIHPDGEKVAQRIRGATDLPKIPIPSVPTFQPS
□ Pp_26_365	38.8	9.4	0.92	SGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVDEATWERMWKHVAKIHPDGEKVAQRIRGATDLPKIPIPSVPTFQPS

for vector NKI_1_1

Tagged constructs in vector NKI_1_1

	MW	pI	ε	
□ Pp_1_310	36.9	8.4	0.93	MAHHHHHHSAALEVLFQGPGMPGGKKVAGGGSSGATPTSAAATAPSGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVD
□ Pp_1_365	43.2	9.4	0.83	MAHHHHHHSAALEVLFQGPGMPGGKKVAGGGSSGATPTSAAATAPSGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVD
□ Pp_26_310	34.8	7.3	0.98	MAHHHHHHSAALEVLFQGPGSGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVDEATWERMWKHVAKIHPDGEKVAQRI
□ Pp_26_365	41	9.3	0.87	MAHHHHHHSAALEVLFQGPGSGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVDEATWERMWKHVAKIHPDGEKVAQRI

Cleaved constructs with 3C

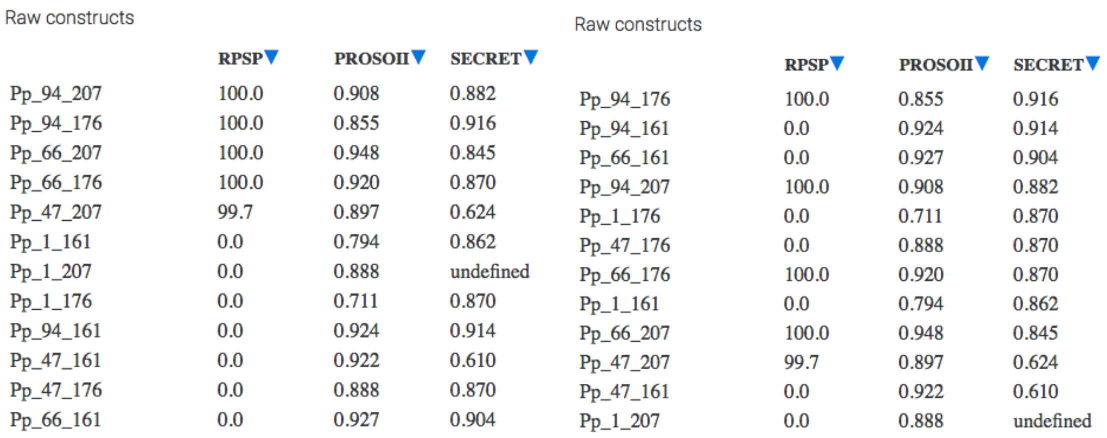
	MW	pI	ε	
□ Pp_1_310	34.9	8.7	0.98	GPGMPGGKKVAGGGSSGATPTSAAATAPSGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVDEATWERMWKHVAKIHPD
□ Pp_1_365	41.2	9.5	0.87	GPGMPGGKKVAGGGSSGATPTSAAATAPSGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVDEATWERMWKHVAKIHPD
□ Pp_26_310	32.8	7.9	1.04	GPGSGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVDEATWERMWKHVAKIHPDGEKVAQRIRGATDLPKIPIPSVPTF
□ Pp_26_365	39.1	9.4	0.91	GPGSGVRRLETSEGTSAQRDEEPEEEGEEDLRDGGVPFFVNRGGLPVDEATWERMWKHVAKIHPDGEKVAQRIRGATDLPKIPIPSVPTF

Figure 4. The three types of constructs, native, tagged and cleaved, are displayed along with key physicochemical parameters.

3.2.4 Scoring of solubility and crystallisability index of proteins and ranking

To score how likely the designed constructs are to be soluble and crystallise we use the RPSP (Harrison and Bagajewicz, 2015), PROSOII (Smialowski et al., 2012) and SECRET (Smialowski et al., 2006) servers. The user can select all or specific constructs and request the scoring. As these predictions take several second and sometimes a few minutes, parallelisation of that task was important. We use the “Celery” technology to allow parallel processing of server requests and handling the outcome of these requests. This allows tracking all jobs, speeding up the response time to the user and efficient feedback as the results come back. The results are returned to the user and display in the CCD.

The peptides chosen by the user chooses are ranked by their RPSP score by default, but users can easily choose a different ranking interactively. An example is illustrated in Figure 5.



Raw constructs

	RPSP▼	PROSOII▼	SECRET▼
Pp_94_207	100.0	0.908	0.882
Pp_94_176	100.0	0.855	0.916
Pp_66_207	100.0	0.948	0.845
Pp_66_176	100.0	0.920	0.870
Pp_47_207	99.7	0.897	0.624
Pp_1_161	0.0	0.794	0.862
Pp_1_207	0.0	0.888	undefined
Pp_1_176	0.0	0.711	0.870
Pp_94_161	0.0	0.924	0.914
Pp_47_161	0.0	0.922	0.610
Pp_47_176	0.0	0.888	0.870
Pp_66_161	0.0	0.927	0.904

Raw constructs

	RPSP▼	PROSOII▼	SECRET▼
Pp_94_176	100.0	0.855	0.916
Pp_94_161	0.0	0.924	0.914
Pp_66_161	0.0	0.927	0.904
Pp_94_207	100.0	0.908	0.882
Pp_1_176	0.0	0.711	0.870
Pp_47_176	0.0	0.888	0.870
Pp_66_176	100.0	0.920	0.870
Pp_1_161	0.0	0.794	0.862
Pp_66_207	100.0	0.948	0.845
Pp_47_207	99.7	0.897	0.624
Pp_47_161	0.0	0.922	0.610
Pp_1_207	0.0	0.888	undefined

Figure 5. Scoring and ranking (left by solubility, right be crystallisability) of designed protein constructs

References cited

Celie, P.H., Parret, A.H., and Perrakis, A. (2016). Recombinant cloning strategies for protein expression. Curr Opin Struct Biol *38*, 145–154.

Harrison, R.G., and Bagajewicz, M.J. (2015). Predicting the solubility of recombinant proteins in Escherichia coli. Methods Mol Biol *1258*, 403–408.

Luna-Vargas, M.P.A., Christodoulou, E., Alfieri, A., van Dijk, W.J., Stadnik, M., Hibbert, R.G., Sahtoe, D.D., Clerici, M., Marco, V.D., Littler, D., et al. (2011). Enabling high-throughput ligation-independent cloning and protein expression for the family of ubiquitin specific proteases. Journal of Structural Biology *175*, 113–119.

Mooij, W., Mitsiki, E., and Perrakis, A. (2009). ProteinCCD: enabling the design of protein truncation constructs for expression and crystallization experiments. Nucleic Acids Res.

Smialowski, P., Doose, G., Torkler, P., Kaufmann, S., and Frishman, D. (2012). PROSO II--a new method for protein solubility prediction. Febs J. *279*, 2192–2200.

Smialowski, P., Schmidt, T., Cox, J., Kirschner, A., and Frishman, D. (2006). Will my protein crystallize? A sequence-based predictor. Proteins *62*, 343–355.