SMP: Protein

Project: Systematic Expression of Human Proteins for Structure Analysis, Antibody Production and Protein Arrays

Konrad Büssow - **Max Planck Institute for Molecular Genetics, Berlin - buessow@molgen.mpg.de**

Introduction

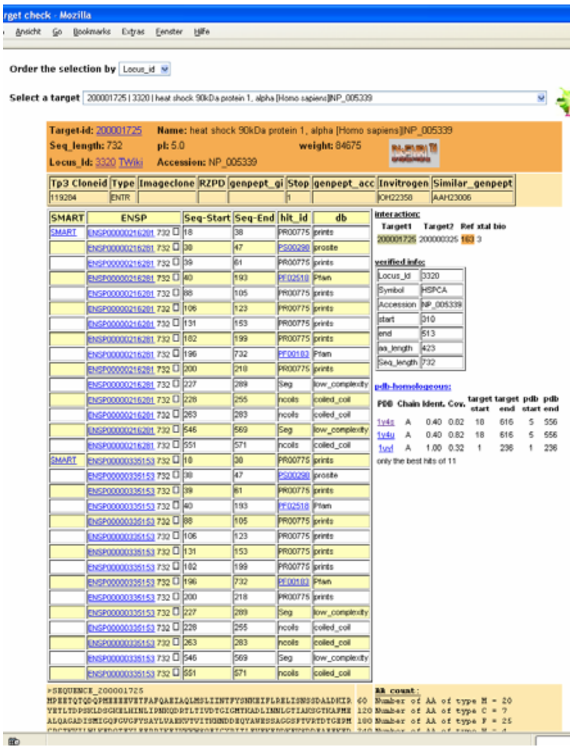
Our project provides human recombinant proteins to our partners in the SMP Protein and to collaboration partners in the NGFN. Proteins are expressed in bacterial and insect cell systems, purified via chromatography and characterised by biophysical assays and optical spectroscopy. We rely on the methods and know-how developed during NGFN-1 in the protein platform and in the BMBF *Leitprojekt* Protein Structure Factory [1, 2]. The Protein Structure Factory (PSF) is part of the international structural genomics initiative [3] and aims at the determination of human protein structures by X-ray diffraction methods and NMR spectroscopy using standardised high-throughput procedures. A complete pipeline has been established for this purpose that comprises cloning, protein expression in small and large scale, biophysical protein characterisation, crystallisation, X-ray diffraction and structure calculation. 537 different human proteins have been studied by the PSF. 139 proteins (18%) could be expressed and purified in soluble form [4]. The structures of ten full length proteins have been solved by crystallography and several domains were solved by NMR (Table 1).

***Tab 1:** The structures of full length proteins solved by the Protein Structure Factory.*

Name	GenBank accession	PDB ID
Gankyrin	AAH11960	1QYM
APEG1, aortic preferentially expressed protein 1	AAH06346	1U2H
Fumarylacetoacetate hydrolase family member FLJ36880	CAB66654	1SAW
PTD012	CAB66540	1XCR
14.5 kDa translational inhibitor protein, p14.5	CAA64670	1ONI
BET3, trafficking protein particle subunit	AAB96936	1SZ7
Peptidase D	AAH28295	
CutC copper transporter homolog, CGI-32	AAH21105	
TPC6	CAI46185	2BJN
Nicotinamide mononucleotide adenylyltransferase	NP_003012	1GZU
FAF2 domain of p47		1SS6
BAG: domain of silencer of death		1M7K
EXOS: domain of Q933J5		1R4T
NADH-ubiquinone oxidoreductase subunit CI-B8		1S3A

Target proteins

The SMP Protein is focussed on protein-protein interactions, which are studied by a variety of methods. The structural biology project of the platform aims at understanding the mechanism of protein-protein interaction at the atomic level. We will solve structures of protein complexes identified by our partners in the SMP Protein. Interacting proteins are produced recombinantly in the expression systems *E. coli* and insect cells/baculovirus. Standard affinity peptide tags and specific protease cleavage sites allow to purify proteins by standard affinity chromatography and to remove affinity tags by proteolysis prior to crystallisation. Protein complexes are assembled *in vitro* or by co-expression (see below). Protein-protein interactions are verified by gel filtration and dynamic light scattering. Isothermal titration calorimetry is performed in collaboration with Prof. K.P. Hofmann, Charité Berlin.



Order the selection by Locus_id

Select a target 200001725 | 3320 | heat shock 90kDa protein 1, alpha [Homo sapiens]NP_005339

Target-id: 200001725 **Name:** heat shock 90kDa protein 1, alpha [Homo sapiens]NP_005339
Seq_length: 732 **pI:** 5.0 **weight:** 84675
Locus_Id: 3320 TWiki **Accession:** NP_005339

Tp3 Cloneid	Type	Imageclone	RZPD	genpept_gi	Stop	genpept_acc	Invitrogen	Similar_genpept
119204	ENTR				1		IOH22358	AAH23006

SMART	ENSP	Seq-Start	Seq-End	hit_id	db
SMART	ENSP00000216281 732 □	18	38	PR00775	prints
	ENSP00000216281 732 □	38	47	PS00298	prosite
	ENSP00000216281 732 □	39	61	PR00775	prints
	ENSP00000216281 732 □	40	193	PF02518	Pfam
	ENSP00000216281 732 □	88	105	PR00775	prints
	ENSP00000216281 732 □	106	123	PR00775	prints
	ENSP00000216281 732 □	131	153	PR00775	prints
	ENSP00000216281 732 □	182	199	PR00775	prints
	ENSP00000216281 732 □	196	732	PF00183	Pfam
	ENSP00000216281 732 □	200	218	PR00775	prints
	ENSP00000216281 732 □	227	289	Seg	low_complexity
	ENSP00000216281 732 □	228	255	ncoils	coiled_coil
	ENSP00000216281 732 □	263	283	ncoils	coiled_coil
	ENSP00000216281 732 □	546	569	Seg	low_complexity
	ENSP00000216281 732 □	551	571	ncoils	coiled_coil
SMART	ENSP00000335153 732 □	18	38	PR00775	prints
	ENSP00000335153 732 □	38	47	PS00298	prosite
	ENSP00000335153 732 □	39	61	PR00775	prints
	ENSP00000335153 732 □	40	193	PF02518	Pfam
	ENSP00000335153 732 □	88	105	PR00775	prints
	ENSP00000335153 732 □	106	123	PR00775	prints
	ENSP00000335153 732 □	131	153	PR00775	prints
	ENSP00000335153 732 □	182	199	PR00775	prints
	ENSP00000335153 732 □	196	732	PF00183	Pfam
	ENSP00000335153 732 □	200	218	PR00775	prints
	ENSP00000335153 732 □	227	289	Seg	low_complexity
	ENSP00000335153 732 □	228	255	ncoils	coiled_coil
	ENSP00000335153 732 □	263	283	ncoils	coiled_coil
	ENSP00000335153 732 □	546	569	Seg	low_complexity
	ENSP00000335153 732 □	551	571	ncoils	coiled_coil

interaction:

Target1	Target2	Ref	xtal	bio
200001725	200000025	163	3	

verified info:

Locus_Id	3320
Symbol	HSPCA
Accession	NP_005339
start	310
end	513
aa_length	423
Seq_length	732

pdb-homologeous:

PDB	Chain	Ident.	Cov.	target start	target end	pdb start	pdb end
1y4s	A	0.40	0.82	18	616	5	556
1y4u	A	0.40	0.82	18	616	5	556
1yvl	A	1.00	0.32	1	236	1	236

only the best hits of 11

```
>SEQUENCE_200001725
MPEETQTQDQPMEEEEVETFAFQAEIAQLMSLIINTFYSNKEIFLRELISNSSDALDKIR  60
YETLTDPSKLDSGKELHINLIPNKQDRTLTIVDTGIGMTKADLINNLGTIAKSGTKAFME 120
ALQAGADISMIGQFGVGFYSAYLVAEKVTVITKHNDDEQYAWESSAGGSFTVRTDTGEPM 180
```

```
AA count:
Number of AA of type M = 20
Number of AA of type C = 7
Number of AA of type F = 25
```

***Fig 1:** The NGFN target check web tool, created by Ulf Lenski, integrates sequence information and links to various analysis tools.*

Protein production for structural biology

The full-length protein is not always the ideal starting point for crystallisation experiments. Many eukaryotic proteins are large and contain several domains. Some also contain sequence stretches which are not folded in the cell but adopt a random, flexible conformation. Protein crystallisation is often complicated by the presence of such sequences. Consider the determination of the APEG-1 protein structure as an example. APEG-1, aortic preferentially expressed gene, is a marker for vascular smooth muscle cell differentiation. Although the protein is small and can be expressed well in *E. coli* cells, no crystals of sufficient quality could be obtained. Sequence analysis revealed that the N-terminal 16 amino acids have a high probability of lacking an ordered structure. A new expression construct was designed without the N-terminal residues. The resulting protein could be crystallised and the structure solved (PDB accession 1U2H) [5].

Based on sequence analysis, we design up to ten expression constructs for a given target protein to increase the chances of successful protein crystallisation. Each target protein is analysed by a variety of different sequence analysis tools (Table 2). The domain structure is determined with SMART and detailed domain information is obtained from PFAM. Functional annotation and protein-protein interaction data is retrieved from OMIM, the NCBI Genes database, KEGG, iPATH and STRING. All target sequences are compared to sequences of known structures from the Protein Data Bank (PDB) with BLAST. Disordered regions are predicted with DISOPRED2 and DISEMBL. Splice variants for target genes are obtained from the ENSEMBL database. Fully sequenced cDNA clones as templates for cloning experiments are obtained from the German Resource Center (RZPD).

An interactive web interface (*NGFN target check*) was created that displays sequence and clone information and annotations and allows performing all relevant sequence analyses easily by following links in a web browser (Fig. 1).

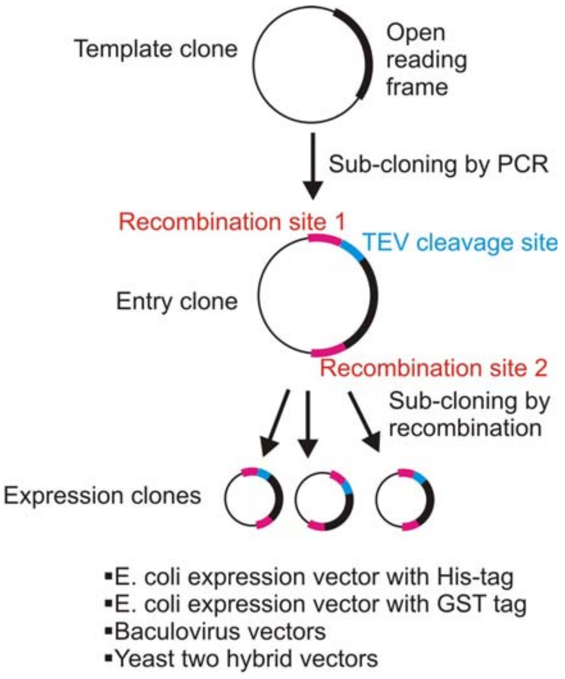


Fig. 2: *The cloning strategy allows to shuttle expression constructs into vectors for a variety of applications.*

Cloning Strategy

The Gateway cloning system of Invitrogen allows shuttling a cDNA expression construct into a variety of expression vector by a standardised recombination step. This avoids the risk of sequence errors introduced by PCR amplification and standard cloning procedures. First, an entry clone is created by standard cloning (Fig. 2). A cleavage site of tobacco etch virus (TEV) protease is included into the constructs, which allows to remove any N-terminal affinity tags after protein expression and purification. When the entry clone has been created and is proven to be free of errors by DNA sequencing, it is used to prepare a set of expression clones by simple recombination reactions. Thereby, the same expression constructs can be studied with a variety of experimental approaches. For protein production, the optimal expression system can be quickly determined.

All clones and protein expression experiments are stored in the project database which is connected to the central database of the Protein Structure Factory. The central PSF database is accessible via web interfaces and integrates data from all experimental steps, from cloning to final structures.

Co-Expression

We have developed a set of novel expression vectors for protein co-expression in *E. coli*. Two or more different proteins can be expressed in the same cell. The vectors have His- or GST-affinity tags and are compatible to the Gateway cloning system. An example is shown in Figure 3.

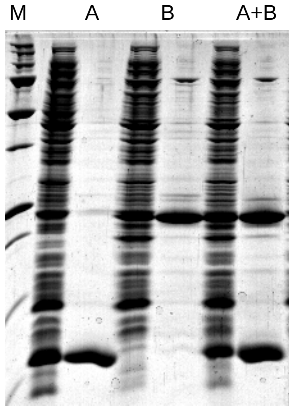


Fig. 3: *Protein co-expression. Crude cell lysates and purified proteins are shown for protein A, protein B and a clone expressing A+B.*

Conclusion

High throughput methods for cloning, protein production and crystallisation are applied systematically for the study of protein-protein interactions in the SMP Protein. The close collaboration between the platform partners and the careful selection of target proteins and complexes ensures a high scientific value of structures solved by the platform.

Table 2: *Sequence analysis tools*

DISEMBL – http://dis.embl.de
DISOPRED2 – http://bioinf.cs.ucl.ac.uk/disopred
ENSEMBL – http://www.ensembl.org
iPath – http://escience.invitrogen.com/ipath
KEGG – http://www.genome.ad.jp/kegg
OMIM – http://www.ncbi.nlm.nih.gov/Omim
PDB – http://www.pdb.org
PFAM – http://www.sanger.ac.uk/Software/Pfam
SMART – http://smart.embl-heidelberg.de
STRING – http://string.embl.de

Lit.: ***1.*** *http://www.proteinstrukturfabrik.de.* ***2.*** *Heinemann, U., et al., Facilities and methods for the high-throughput crystal structure analysis of human proteins. Acc. Chem. Res., 2003. 36(3): p. 157-163.* ***3.*** *Zhang, C., et al., Overview of structural genomics: from structure to function. Curr. Opin. Chem. Biol., 2003. 7(1): p. 28-32.* ***4.*** *Büssow, K., et al., Structural Genomics of human proteins - target selection and generation of a public catalog of expression clones. Microbial Cell Factories, 2005. 4: p. 21.* ***5.*** *Manjasetty, B.A., et al., X-ray structure of engineered human Aortic Preferentially Expressed Protein-1 (APEG-1). submitted for publication, 2005.*