P4EU workshop

Construct design for recombinant protein production

Friday 15 December 2023, 9:30-11:30 CET

Please name the topic here

Please type your question(s) to the topic here.

PLEASE KEEP THIS TEMPLATE SLIDE!

Copy/paste & edit to add new topics.

THANKS

Domain boundaries for expression constructs

How to define suitable domain boundaries? Should I make the full protein or just a domain?

What are the end amino acids to be avoided? Also, the other way round, which aa to start with best?

Replies:

-AF2/PDB homologous models

Linker sequences between target and tag

What is your favourite linker? GGS? GST? What number thereof?

How short or long should a linker be?

What is the minimal linker length to consider the two domains independent?

Which linker should one use for which purpose - rigid, flexible linkers and so on

Alena prefers: GSAGSAAGSG

Aljaz: 2xGSS

Anja: G4S linker

KL:Native linker if present

KR: nice paper on linkers is doi:10.1016/j.addr.2012.09.039

-Optimal linker: any studies done?

Secretion & signal sequences

When you secrete a human protein from HEK or CHO cells, do you prefer to keep the native signal sequence or do you generally replace the native signal sequence with a generic one such as e.g. IL2? If you use a generic one, what is your favorite signal sequence?

Would you consider to secrete an natively intracellular protein to the cell culture medium by placing non-native/artificial secretion signals?

Where would you place the secretion signal?

How do you know that the secretion signal is correctly cleaved?

Secretion & signal sequences

Replies:

NB: mu phosphatase (eukaryotic) - **MTSTLPFSPQVSTPRSKFILNSYNQRRYTMGILPSPGMPALLSLVSLLSVLLMGCVA** ! *ETG*

AB: Honey bee melittin (Insect) and start with endogenous. Secreting intracellular proteins not particular efficient, but may work

SAM: project dependent, might change expression levels for complexes. If not native, uses mouse IgG light chain in HEK

SS: Native and also non native if you have capacity. Compare also with SignalP as relative to the N-terminus of protein. Check cleavage by intact mass if possible. Secretion of non-secreted proteins: Tsafi presented data that said it was not so easy

LS: Intracellular outside, prefer to do intracellular in insect. Keep in mind PTM, and it's a totally different environment, reducing vs oxidizing. Spike SS, C-terminal His-tag

KR: For secreted proteins, tag in C? N-terminal may interfere with cleavage also due to charged His

AB: pelB for nanobodies
KR: if not secreted to periplasm, use SHuffle for nanobodies. Some nanobodies also stable even without formation of the SS-bonds.

SS: How to detect disulfide by MS - by intact mass; AG: or by MS/MS

Fusion tags & solubility

What is your favorite solubility-enhancing fusion tag? Does your choice vary depending on the expression system (*E. coli*, yeast, insect or mammalian cells) that you use?

Is there a combination of affinity-solubility tag that should be avoided?

Should an AlphaFold prediction be performed on the final Tag-Protein sequence?

Position of the solubility tag- what is better- N-terminal or C-terminal fusion ?

KR: normally you place a solubility-enhancing tag at the N-terminus, but e.g. for IDPs it can be useful at the C-terminus as well
YP: bdSUMO (not yeast!) at N-terminus; favourite. Cleavage is efficient and leaves no non-native amino acid overhang
NB: after cleavage, more likely to have soluble proteins. With MBP and and GST sometimes have false positives (falsely soluble)
NB: Sumo>Trx=Z>>>mbp>GST, in eukaryotic : mostly FP
LS:SUMO, GST/MBP
IS:MBP binds sometimes unspecific, she only uses for things not for cleavage
SS: MBP folding helper/chaperone, but in reality it’s just a soluble aggregate. pI of 5 and if your target its basic, you never get them separated.

Codon optimization

Would you generally recommend to use a codon optimisation algorithm when ordering a synthetic gene? Would you worry about the specific effect of rare codons on translation speed and how that relates to protein folding and possibly disturbing that relationship during codon optimisation? What is your favorite codon optimisation algorithm?

Which tool/website would you recommend to quickly and easily check the codon usage of a sequence and adjust it if necessary?

AB : in general, if ordering synthetic, do codon optimization. Full vs slight opti : slight was better, balance between codons and secondary structure
CC:10.1016/j.pep.2008.01.008 Sometimes suppliers optimize for synthesis and not necessarily expression
KL: optimizing codons or secondary structures.
LS: optimized sequence nothing, but native sequence worked well. Protein dependent, in general opti, but sometimes could be worst. This was also a human sequence.
SAM: once back to back, HEK secretion, opti with native secretion sequence, 20% more. This was a human protein.
SAM: asks Genscript algo
KR: Geneart algo
AB: Genscript. Also optimized sequence might interfere with UTR
SK: how much gain in optimization?
KR: hard to get nice data. How about human protein in HEK? What to do? Tend to not do that
YP: in Ecoli, typically not that big of a problem, its when they are in tandem

Choose suitable expression system

What expression system (host) should I choose to produce my protein of interest?

(I have a vague idea that there is a nice recent publication on this subject ;-)

Up to which protein size can be expressed in E.coli and are there any special aspects that need to be considered when expressing large constructs (>150kDa)?

Down to which protein size can be expressed in E.coli and are there any special aspects that need to be considered when expressing/designing small constructs? Does this also is related to if it is structured or not?

KR: New P4EU paper with a workflow to decide: https://star-protocols.cell.com/protocols/3094
Size limit is protein dependent
YP: 300 ish kDa, TB, Fatty acid synthase and works well
AB: Dynein, 250 kDa but insoluble
SS: 260 kDa in Ecoli, structure dependent

IDP and ID domains

1. Does anyone have any experience on expressing IDPs and/or designing constructs with ID domains?
2. Has anyone tried treating IDPs as membraned proteins in Cell-Free systems? i.e. can you satisfy IDP protein requirements with additives/environments similar to IMPs?
3. How do you go about to try to figure out if a part of a protein that was defined as unstructured in alphafold is in fact an ID region or a stretch of protein that alphafold was not able to find the structured domain.

BB: High Salt/Arginine. ‘Loose resin’ or else precipitates. Difference in lysis. High pressure no sonication. Only fresh pellet. Keep in mind nucleic acids need to be sheared.
AB: Keep MBP until the end, or phase separate. Salt minimum 0.5 M. In Insect cells. Do not try Ecoli anymore
CC: IDP, Secreted MBP in HEK.
CC: Anyone ever use 1,6-Hexanediol? AB, could separate different species with this by SEC

Expression of protein complexes

What is the most efficient strategy to express protein complexes? Polycistronic vectors or a simple co-transfection? Do you have any experiences with auto-cleavable peptides (e.g. efficiency of the cleavage?) Where do you place the tag?

KL: fo Ab coinfection works well. But you start to see differences in stables and more sensitive higher-order complexes (>3-4 genes)

AB: multicassette vector, instead of co-transfection. Especially if the subunits rely on each other to fold.

LS: Also on one vector. Where tag: placing the tag on lowest abundance allows to purify out other subunits in excess

AB: Tag on one that is most exposed. Sometimes the complexes, separated into subcomplexes

SS: Multibac, BigBAC,GoldenBAC. For HEK?
JVDH: BacMam

YP: Position of tag very empirical

JVDH: Single vectors, much easier to manipulate on their own

NB: P2A??
AB: 50% +/- efficient. Need to separate the fusion and the components. Not 100% in HEK

Expression of protein complexes

Replies:

SAM: cotransfection with 2 plasmids, and tagged on Beta subunit for integrins.

JM: Use IRES, for two proteins.

SS/JVDH: for mammalian, vector size limit, 30 kbp. Transfection is still limited. BacMam supports more.

KR: 4 vectors by transient in ExpiHEK
JVDH: is it coexpression in ONE cell, or assembles during extraction?

LS: Purification from endogenous sources. Tagged subunit
KR: CRISPR in tagged subunit. But tends to loose weak interactors and complex dependent

OV: From Imre: Long polypeptide with protease cleavage sites Nie Y, Bellon-Echeverria I, Trowitzsch S, Bieniossek C, Berger I. Multiprotein complex production in insect cells by using polyproteins. Methods Mol Biol. 2014;1091:131-41. doi: 10.1007/978-1-62703-691-7_8. PMID: 24203328

Facilitator DNA elements

Do you have any experience with adding facilitator DNA elements between enhancer and promoter regions on your expression vectors with regards to yields of secreted and intracellular proteins?

KL: tried for Ab. Differences betw vectors sets depending on the element.

Proteasome degradation

Any experience with degradation of expressed target, how to predict target sequences that should be mutated in order to abrogate proteasome degradation in cell?

> AB: Never had such a problem. Look in literature and homologous ones to see if they are targeted for degradation

Forcing post-translational modifications onto proteins

Any experience (during expression) with removal of phosphorylation, forcing site-specific phosphorylation, mono-ubiquitylation etc?

Any success with others?

SK: for phosphorylation, coinfect/transfect with the required kinase. Can also pull out interacting partners

LS: put in a glycosylation site between two domains to break a dimer

JVDH: for phospho, site directed mutagenesis to Asp/Ala.

Summary (AS)

A: General Considerations

1. **Consider Protein Domain Boundaries**
    - Use secondary structure predictions to define domain boundaries more precisely. Knowing the locations of alpha helices, beta strands, and other structural elements can aid in selecting appropriate boundaries for expression constructs.
    - Use available structural and functional information.
    - Design constructs that encompass complete domains & secondary structures to maintain structural integrity and functional activity.
2. **Protein Size**
    - Be mindful of the protein size. Extremely large proteins may be challenging to express and purify. Consider breaking large proteins into functional domains for easier handling.
3. **Avoid Unstructured Regions**
    - Exclude highly flexible and unstructured regions that are prone to aggregation and degradation.
    - Focus on regions with defined secondary structures to enhance protein stability.
4. **Mind Signal Peptides and Transit Sequences**
    - If the protein has a signal peptide or transit sequence for cellular targeting, include it in the construct for proper localization.
    - Consider removing signal peptides if the goal is to express the mature form of the protein.
5. **Avoid Transmembrane Domains (if applicable)**
    - Exclude transmembrane domains if the aim is to express a soluble protein.
    - Consider alternative constructs or fusion partners for membrane protein expression.
6. **Optimize Codon Usage**
    - Optimize codon usage for the expression host to enhance translation efficiency.
    - Avoid rare codons that may lead to translational pauses and reduce protein yield.

Summary (AS)

7. **Incorporate Fusion Tags Wisely**
    - Introduce fusion tags for purification, detection, or solubility enhancement. Common tags include His-tag, GST-tag, or MBP-tag.
    - Choose fusion tags compatible with downstream applications, and consider their potential impact on protein folding and function.
    - Tag placement: Check if placing the tag at either the N- or C-terminus may interfere with function. Alternatively, placing tags in regions predicted to have loops or turns can minimize interference with structured regions.
8. **Check for Post-Translational Modifications**
    - Consider the presence of post-translational modification sites (e.g., glycosylation, phosphorylation) and include them if necessary for proper protein function.
9. **Maintain Optimal Size for Expression**
    - Avoid excessively large constructs that may lead to low expression levels, inefficient translation, or difficulties in downstream processing.
    - Optimize construct size for the specific expression system being used.
10. **Test Iteratively**
    - Design multiple constructs with different boundaries to test and identify the most successful one for expression.
    - Consider using truncation mutants to pinpoint regions essential for expression and function.
11. **Consider Fusion Proteins for Difficult Targets**
    - For challenging proteins, consider fusing the target protein with a well-expressed and stable partner to enhance expression and solubility.

Summary (AS)

12. **Verify Structural Integrity**
    - Validate the structural integrity of the designed constructs through bioinformatics tools, such as molecular dynamics simulations or structural prediction methods.
13. **Use Compatible Vectors and Host Systems**
    - Choose expression vectors that are compatible with the selected host systems.
    - Choose the expression system (bacterial, yeast, mammalian cells, etc.) based on the requirements of your protein. Different systems have distinct post-translational modification capabilities and folding environments.

By following these guidelines, you can design protein constructs that maximize expression levels while preserving the structural and functional characteristics of the target protein. Iterative testing (expression construct screening) and optimization are crucial steps in the process to ensure the successful production of the desired protein.

Summary (AS)

B: Domain Swapping

1. **Domain Boundaries**
    - Clearly define domain boundaries based on structural and functional information. Consider using flexible linkers to maintain the independence of domains.
2. **Linker Optimization**
    - Optimize linker length and composition. Too short linkers may lead to steric clashes, while too long linkers might reduce the stability of the fusion protein.
3. **Validation**
    - Validate domain swapping through structural and functional assays. Ensure that the swapped domains retain their native conformations and functions.

Summary (AS)

C: Linker Design

1. **Flexible vs. Rigid Linkers**
    - Choose linkers based on the desired flexibility. For rigid connections, short and structured linkers may be preferred, while longer flexible linkers can be used to provide more freedom of movement.
2. **Composition**
    - Consider using glycine-serine-rich linkers for flexibility. Avoid proline-rich linkers, as they introduce kinks and rigidity.
    - On the other hand, the rigidity introduced by proline can help maintain the spatial separation of domains, especially in multi-domain or multi-subunit proteins.
    - Flexible linkers may be necessary for connecting domains with different secondary structure elements.
3. **Linker Length**
    - Optimize linker length empirically. This may involve trying different lengths to find the optimal balance between flexibility and maintaining the structural integrity of the linked domains.

Summary (AS)

D: Non-Structured Regions

1. **Intrinsically Disordered Regions (IDRs):**
    - If your protein contains intrinsically disordered regions, consider their role in function and structure. You may need to experiment with truncations or stabilizing strategies.
2. **Stabilization Techniques:**
    - Implement strategies such as the introduction of stabilizing mutations or fusion to a stable partner to improve expression and structural determination.
3. **Biophysical Characterization:**
    - Use complementary biophysical techniques like circular dichroism (CD) or nuclear magnetic resonance (NMR) to characterize non-structured regions and understand their dynamics.

Summary (AS)

E: Pro-rich regions

1. **Disruption of Secondary Structure**
    - Proline is known to disrupt regular secondary structures such as alpha helices. If the protein requires specific secondary structures in certain regions, proline-rich regions may interfere with the desired conformation.
2. **Potential for Misfolding**
    - While proline can stabilize certain structures, in other contexts, it might lead to misfolding or aggregation. This is particularly important if the protein is prone to aggregation or if the proline-rich region interferes with proper folding.

3. **Impact on Solubility**
    - Proline-rich regions may affect the solubility of the protein. If high solubility is crucial for downstream applications, such as structural studies or functional assays, the inclusion of proline-rich regions should be carefully considered.

Summary (AS)

F: Avoid aggregation-prone regions

Aggregation-prone regions in proteins refer to sequences or structural motifs that have a higher propensity to self-associate and form aggregates. Protein aggregation can lead to the formation of insoluble aggregates, inclusion bodies, or amyloid fibrils, which can be detrimental for protein expression, purification, and functional studies. Identifying and addressing aggregation-prone regions is crucial in the design of protein expression constructs. Here are some characteristics of aggregation-prone regions.

1. **Hydrophobicity**
    - High hydrophobicity is a common characteristic of aggregation-prone regions. Exposed hydrophobic residues have a tendency to interact with each other, leading to protein aggregation.
2. **Beta-Sheet Rich Sequences**
    - Regions with a high propensity to form beta-sheet structures are often associated with protein aggregation. Beta-sheet interactions can contribute to the formation of aggregates and amyloid fibrils.

3. **Amyloidogenic Sequences**
    - Amyloidogenic sequences, which have the potential to form amyloid structures, are highly aggregation-prone. These sequences often contain short, repetitive motifs with a tendency to stack and form beta-sheet structures.
    - Amyloidogenic sequences often involve amino acids with a high propensity to form beta-sheet structures, such as beta-branched residues (valine, isoleucine), glycine, serine, and aromatic residues (tyrosine, phenylalanine).
    - Polar and charged residues may also be involved in the formation of amyloidogenic structures.

Summary (AS)

4. **Low Complexity Sequences**
    - Low complexity regions, characterized by the repetition of specific amino acid residues, can be prone to aggregation. These regions may lack defined structures and promote self-association.

2. **Polyglutamine Tracts**
    - Proteins containing polyglutamine tracts are known to be aggregation-prone. Expansion of polyglutamine repeats is associated with several neurodegenerative diseases.

3. **Proline-Rich Regions**
    - While proline is often considered a helix breaker, proline-rich regions can contribute to aggregation, especially if they form extended conformations or beta-sheet structures.

4. **Unstructured or Disordered Regions**
    - Intrinsically disordered regions or unstructured loops may lack defined structures, facilitating interactions that lead to aggregation.

5. **Exposed Aromatic Residues**
    - Exposed aromatic residues, such as tyrosine and phenylalanine, can contribute to aggregation due to pi-pi stacking interactions.

Summary (AS)

9. **Length of Unstructured Loops**
    - Long and flexible unstructured loops, especially those connecting structured domains, may be prone to aggregation.

2. **Net Charge**
    - Regions with extreme net charge (either highly positively or negatively charged) may be aggregation-prone, as charge-charge interactions can drive self-association.

3. **Context-Dependent Factors**
    - The aggregation propensity can also depend on the context of the surrounding sequence and the protein's overall fold. A seemingly innocuous region in one protein may become aggregation-prone in a different protein context.

4. **Predictive Tools**
    - Various bioinformatics tools and algorithms, such as TANGO, Waltz, and PASTA, can predict aggregation-prone regions based on sequence features.

5. **Experimental Validation**
    - Experimental techniques like dynamic light scattering, size exclusion chromatography, and Thioflavin T assays can be used to experimentally validate the aggregation propensity of specific regions.