Forum for European
Structural Proteomics

Research Infrastructures in Structural Proteomics: Assessment of Needs for Protein Production in Structural Genomics / Structural Proteomics Projects.

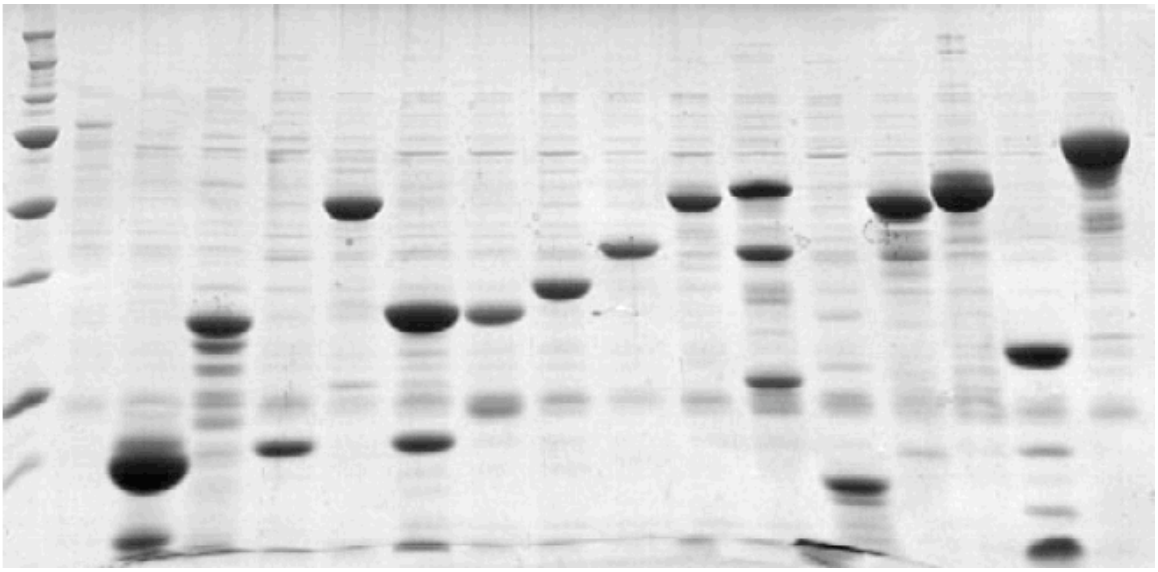


Table of contents

Executive Summary

Proteins are central players in all processes of life. They may act as cellular scaffolds and mediate almost all biological processes. Structural Genomics/Structural Proteomics (SP/SG) projects aim at understanding protein structure and function at the molecular and atomic level. They differ from classical structural biology (SB) in their systematic approach to protein target selection and in the elevated throughput in structure determination that they are striving to achieve. For protein structure analysis relatively large amounts of pure individual proteins or protein complexes are required; hence, the production of suitable protein samples is a central prerequisite for success. Despite recent advances in technology, protein production is commonly viewed as a rate-limiting step in SG/SP projects. This study by FESP (the Forum for European Structural Proteomics) aims at identifying the factors that cause protein production to be rate limiting and at recommending measures to deal with them.

In this context, FESP conducted a survey in which a questionnaire was distributed to structural biologists with group leader status, located primarily at universities and research institutes. A total of 77 scientists completed the questionnaire, and provided valuable input. Their answers showed a profound knowledge of various aspects of protein production, including extraction from native sources, gene cloning, over expression in bacterial and eukaryotic systems, *in vitro* expression, protein purification under native and denaturing conditions, protein refolding, biophysical characterization and measures to overcome limitations on protein solubility (Q11 – Q22)[1]. This pool of experience is, however, very heterogeneous. Especially when it comes to expression of proteins in hosts other than *E. coli,* a bacterium, widely used for recombinant protein production, protein characterization, and protein refolding from insoluble aggregates, the range and prioritization of experimental approaches varies substantially from laboratory to laboratory. While this is in part due to varying technical requirements (*e.g.* sample preparation for X-ray *versus* NMR structure analysis) and to the specific aims of individual projects, it would be desirable to offer specialized hands-on courses organized by and set up for researchers involved in protein production in SG/SP contexts, in order to train scientists and technicians in 'state-of-the-art' methods of protein production for structural analysis. The majority of the contributors to the survey are convinced that such training would improve their productivity (Q28).

The survey also suggests that the limited availability of data concerning protein production experiments carried out in the past, which leads to redundant efforts in many laboratories, further serves to hamper current protein production efforts. This includes, in particular, data regarding expression systems, purification protocols, information concerning failed expression and purification attempts, and which groups have worked on a particular protein and other details (Q29). 71% of the participants stated that they would like to have access to a database storing such information, and would be willing to provide data concerning their own experiments. The average time for entering data should, however, not exceed 30-45 minutes for a particular protein target. An existing database, with correlative functions is the PepcDB of the Protein Structure Initiative (PSI), which is hosted by the NIH. We wish to propose the creation of a similar database for the European SG/SP projects, which might also include individual laboratories, not necessarily doing HTP work. In order to encourage participation, we

[1] The items in the questionnaire are referred by their question number, e.g. Q11.

propose to offer incentives such as privileged access to certain data to laboratories willing to deposit relevant data in such a repository. It is our expectation that such a service will diminish redundancy, enhance exchange of information and collaboration between laboratories, and hence lead to a more efficient use of resources.

Other means of optimization might involve streamlining of laborious and repetitive tasks by standardization of procedures, automation and/or outsourcing: The survey revealed that most groups use bacterial expression for protein production, and use restriction/ligation-based cloning, which in general does not allow fast recloning. It has been shown, however, that expression in eukaryotic hosts and use of solubilizing fusion tags can greatly increase soluble expression of certain proteins, which express as inclusion bodies in bacteria. The use of a unified vector platform allowing access to bacteria, yeast and other eukaryotes, which permits production of protein as fused with solubilizing tags, and which allows fast recloning, would be expected to streamline many or all these aspects of expression testing. Current commercially available vector platforms, which provide these features, do not seem to be optimal for the requirements of X-ray crystallography or other analytical techniques, because they may add additional unstructured peptides to the target protein. Therefore, it may be necessary to modify a commercially available unified vector platform, or to create a new one, in order to satisfy these requirements. In addition, the outsourcing of certain tasks such as the design and evaluation of cloning strategies, gene cloning, generation of cell lines or bacterial strains for recombinant expression, and testing and optimization of protein production, would greatly relieve small groups from the burden of tasks that can readily be automated. In most groups these tasks are performed by group members and, in general, consume a lot of valuable man hours. Only a few of the participants in the survey have used the services of non-profit protein production centers, yet 61% would like to use their services. Presumably, limited access to protein production centers has prevented more widespread use in the past We believe that the support of existing facilities of this kind, and the establishment of new ones (Q30), would greatly increase the work efficiency of many SB laboratories for a relatively modest investment and at a moderate cost.

Introduction

Proteins are central players in all processes of life. They may act as scaffolds for biological cells or be involved in cellular functions such as uptake of nutrients, transport processes, metabolism, sensory perception, signal transduction, cellular interaction, motility, growth and differentiation. In addition, proteins play important roles in all diseases and pathologic processes.

Structural Genomics/Structural Proteomics (SG/SP) aims to provide insight into functions and interactions of proteins at the molecular and atomic level, using complementary experimental techniques and scientific disciplines (X-ray crystallography, NMR spectroscopy, electron microscopy, and biophysics). Protein structure analysis requires large amounts of highly purified proteins and complexes. Hence, the central prerequisites for studies in this field - especially in high-throughput (HTP) environments - are the provision of suitable – usually recombinant – protein sources, the optimization and up-scaling of protein production, as well as the establishment of purification procedures to provide the required protein samples in sufficiently large amounts and of high quality. These activities are collectively denoted as protein production in the following.

In the context of current and future SG/SP projects within Europe, the EC requested the assessment of protein production, which quite frequently has been perceived as a bottleneck in this field. In this assessment, protein production is surveyed along with the large-scale facilities used for crystal, NMR and EM protein structure analysis, although most protein samples are being produced in small or medium laboratory settings. The survey was conducted by FESP (the Forum for European Structural Proteomics, for details see http://www.ec-fesp.org), which, in parallel, assessed the crystallographic, NMR, EM and bioinformatic infrastructures available in Europe. It is anticipated that the data gathered by FESP may aid the EC in developing policies with regard to large infrastructures and SG/SP.

This part of the FESP survey analyzes protein production facilities and methodologies both in current SG/SP projects and in conventional structural biology (SB) laboratories, with an emphasis on identifying rate-limiting steps. Subsequently, suggestions are made as to how to accelerate and economize protein production, and hence to facilitate access of researchers to protein samples. The data presented in this report were acquired using a questionnaire that was sent to the heads of research groups involved in SB all over Europe, from academia as well as from the pharmaceutical industry.

The survey addressed the following questions:

- Which sources of information are generally used, and which considerations guide protein production strategies?
- What is the origin of the protein studied and which expression systems are required for protein production?
- Which methods are applied for protein purification and characterization?
- Which measures are taken to overcome typical problems in protein production and how successful have they proven to be?

- Who is involved in the processes of production and what services could be provided to streamline protein production and save resources?

The findings of the survey are summarized in this report. It provides information about the current status of protein production as a crucial step in all SB and SG/SP being done at universities, research institutes and the pharmaceutical industry throughout Europe. We hope that this survey will provide a frame of reference for shaping policies concerning technologies, infrastructures, and services required for protein production in an SG/SP context.

We thank the participants in our survey for their time and for their willingness to share their individual expertise with respect to many aspects of protein production for SG/SP. The findings and conclusions of this report, including any errors, are however, the full responsibility of FESP.

Protein production in an SP/SG context

The isolation of pure individual proteins or protein complexes is a prerequisite for any kind of macromolecular structure determination. Protein production is hence a central task in all SG/SP work. The introduction of HTP technology in this field has resulted in increased availability of proteins. Despite these advances in technology the production of soluble proteins in the required amounts and with adequate purity, in particular for structure determination using X-ray crystallography and NMR spectroscopy, frequently remains a rate-limiting step for structure analysis. This survey was designed to detect bottlenecks in protein production and to identify measures to overcome them.

Groups contributing data to this survey and their research aims

667 principal investigators (PIs) involved in SB research were selected and invited to participate in the survey. The invitation letters were sent twice to all those on the list, and a further round involving personalized invitations was undertaken to fill perceived gaps in the group of participants. We retrieved 77 valid forms. Not all questions were answered by all participants. One reason may be that in some cases adequate documentation to answer the question may be lacking. Another may be that certain questions do not apply to all labs. We do not perceive this as a serious shortcoming, since no answer may be better than an answer that is not based on solid experience or data. Most questions posed could be answered with “always”, “frequently”, “occasionally” or “never”. In the following, we have combined the first two of these categories as “common” in order to simplify evaluation of the questionnaire.

The responses to the survey came from research groups from 19 European countries. The largest contribution was from Germany, followed by the United Kingdom, Italy, Sweden and Greece (P001). The average time of service as a PI varied greatly, from 1 to 40 years, with a median of 10 years (P002). Most participants groups were from universities and public research institutes (53% and 42%, respectively). The contribution of SBs from industry (2%) and government institutes (2%) was considerably smaller (P003). Most receive national funding (95%), followed by funds from the EU (38%),

private foundations (33%), international support (21%), industry (16%), and the US government (7%) (P004).

The number of employees varies considerably within the groups directed by the PIs. The median numbers are: 3 postdoctoral fellows or research assistants (P005), 3 graduate students, and 2 support staff.

The predominant experimental approaches used by the participants in this study are X-ray crystallography (88%) and biophysics (57%). In addition, NMR spectroscopy (30% and electron microscopy (12%) are used in some laboratories (Q6). These techniques are used to study human proteins and, to a lesser extent, proteins from other eukaryotic organisms (mouse, pig, cow, fruitfly, nematode, plants). Another emphasis is on proteins from *eubacterial* species, including *E. coli*, *B. subtilis*, *P. aeruginosa*, and others. Proteins from *Archaea* are studied less frequently. This is surprising, because *archaeal* proteins have provided valuable insight into the cellular machinery of *eukaryotes,* as certain proteins exist in both kingdoms and display similar functional characteristics. In contrast to the *eukaryotic* proteins, their *archaeal* orthologs quite frequently display a simpler architecture, with fewer compact domains or residues, and hence are better suited to structural studies.

Information retrieval prior to protein production

Information for cloning and expression of proteins was derived from NCBI sources (including PubMed), as well as services by ExPASy, including the SWISS-PROT database, the Protein Data Bank, and Google. Ensembl (a joint project of the EMBL-EBI and the Sanger Centre) a genome browser and SuggestES (developed at the Israel Structural Proteomics Center) a homology-based service for making suggestions for suitable expression systems for a given protein sequence, are used to a lesser extent (Q8).

When a project is initiated, an average of about 10 hours are invested in literature research on a particular protein (Q9). In most cases the proteins to be studied are not obtained from a commercial source (Q10). This may be due either to lack of availability or to high prices demanded for the quantities required for structural studies.

Expression systems are selected mainly by personal experience and, to a lesser extent, on the basis of advice from colleagues and on the scientific literature. Interestingly, only a minority (<20% of all participants) commonly makes use of data available in databases for selecting expression systems for a given protein (Q11). This may be due to the fact that only limited information on expression trials is available in databases, and it is usually not easily accessible. One service, which does exactly this, is SuggestES. But, as just mentioned, it is utilized only sporadically by a minority of the participants of this study. To explain this discrepancy, further inquiries will be necessary. We can only speculate that either this service is not well disseminated within the scientific community, and/or that the data provided are perceived to be inadequate to serve as a reliable suggestion for taking a decision.

Protein production using native sources and recombinant expression

Currently, protein production utilizes either isolation from native sources or recombinant expression using engineered vectors and host systems. The first approach has the advantage that the protein of interest is produced in its biological environment and context; thus, the amounts produced match its functional requirements and do not overload the storage capacity of the cells. In addition, protein folding is usually performed very efficiently by this approach as all native post-translational modifications can occur, and the biologically relevant cofactors are available in the native cells. For most proteins, however, the available amounts are too small to allow purification to homogeneity in amounts, which will permit subsequent structural studies. When it comes to proteins from vertebrates, in particular human proteins, the availability of appropriate cells and tissues may also be limited. These concerns are reflected by the fact that less than half of the participants stated that they purify proteins from native sources, and even within this group, this method is used for <15% of all targets. The proteins obtained from native sources are mainly derived from certain vertebrates (cow, frog, rabbit reticulocyte lysate) and *eubacteria* (*E. coli*, *Pseudomonas* species) (Q14).

Depending on the system used, recombinant expression of genes for protein production excludes one or more of the benefits discussed for isolation from native sources. Its big advantages are, however, the larger yield of the protein of interest and the possibility to introduce affinity tags to facilitate protein purification, heavy-atom labels for crystallographic phasing, or isotopes for NMR experiments. As a consequence, most participants (65) stated that they frequently use this technique, with an average contribution of about 95% to their protein productions. In most groups (> 80%) plasmid-based expression in *E. coli* is commonly used for recombinant protein production (Q14). Various vectors and well established protocols allow fast cloning of genetic constructs and the generation of recombinant strains. Furthermore, *E. coli* cultures grow very fast under standard laboratory conditions, allowing efficient synthesis of recombinant proteins, and cell lysis is straightforward. All these advantages explain why *E. coli* is almost exclusively used for protein production in prokaryotes.

Recombinant *eukaryotic* expression systems are frequently used by <20% of the participants in this survey. They utilize insect cells, yeast and mammalian cell lines.

In addition to protein production in intact organisms, various cell-free expression systems exist. In this survey, only small a subgroup (11 participants) reported the use of these techniques, and even in their laboratories cell-free expression is used rarely, comprising <10% of all protein production trials (Q13). This limited use may be due to the fact that cell-free expression systems require a setup, which is perceived as being expensive, and the necessary reagents need to be obtained commercially. Moreover, upscaling of production is not easy; consequently, synthesis produces a few milligrams per trial at best.

All recombinant expression approaches require the cloning of amplified genetic material (inserts) into genetic vectors. In the early days of molecular biology this was solely done using vectors containing various multiple cloning sites that could be cleaved by restriction enzymes, followed by enzymatic ligation with correspondingly cleaved inserts. This method of cloning is very cost-efficient; however, the standard design of vectors and inserts used in this approach neither allows the cloning of many different genes in parallel nor the efficient switching between vectors which would permit recombinant expression in different organisms or the use of different purification tags.

We were, therefore, surprised to see that the majority of the participants in this survey (>70%) still use these cloning methods (Q12) to a large extent. In contrast, ligase-free T4-polymerase-based cloning and topoisomerase-based cloning are regularly used only by ~10% of all participants. Recombinase-based cloning, for which a large platform of genetic vectors is commercially available, and which allows the quick and efficient cloning and recloning of genes into various vectors, hence overcoming the bottleneck associated with standard restriction/ligation based cloning as described above, is regularly used by only a tiny minority (~5%) of the participants. This may be due to economical considerations, since all competing methods are more expensive than restriction/ligation based cloning. Moreover, the commercially available recombinase-based cloning approaches require sites of recombination which result in additional coding sequences, and hence extend the sequences of cloned genes, resulting in N- and C-terminal extensions in the synthesized proteins which may have negative effects on the subsequent structure determination, *e.g.* on crystallization.

Production of proteins in insoluble fractions and protein refolding

After cell lysis, overproduced recombinant proteins will be found in the soluble and/or the insoluble fraction of the lysate. With the exception of membrane-associated proteins, the presence of the over expressed recombinant protein in an insoluble fraction is an indication that it is not natively folded. If this is the case, variation of growth conditions (*e.g.* medium composition, temperature) or expression parameters (decrease in expression levels, co-expression of ligands), or changing the expression system are all possible approaches to increasing the amount of natively protein in the soluble fraction. The sequestration of overexpressed proteins in the insoluble fraction (*e.g.* as inclusion bodies in bacteria) is sometimes used to obtain large amounts of already relatively pure protein, since quite frequently very few cellular proteins end up in this fraction. In the laboratories represented in this survey, protein purification under denaturing conditions is performed only to a small extent: On average, <15% of all productions involve this approach (Q15). In order to obtain natively folded protein, the insoluble protein fraction needs to be solubilized using denaturants, followed by refolding in an appropriate buffer, a procedure which often requires extensive optimization. Moreover, quite frequently, refolding conditions cannot be found. About half (43) of the participants in this survey state that they use protein refolding at least occasionally. In these laboratories, dialysis-based refolding appears to be the most commonly used method (18%), followed by fast-dilution-based refolding (13%), and on-column refolding (12%) (Q16). Refolding of proteins *in vitro* in the presence of chaperones does not seem to play a significant role (2%). These percentages correlate with the relative success rates of these individual techniques. In most participating laboratories, the outcome of refolding approaches is routinely analyzed by determining protein concentrations in the soluble fraction (43%), and by investigating the hydrodynamic properties of the soluble proteins by gel-filtration chromatography (35%) and light scattering (26%) (Q17). To a lesser extent, the solubilized material is characterized by activity tests, protein binding studies or spectroscopic analyses (NMR and CD). Only few groups (<10%) use cell-based assays, protease resistance, or infrared spectroscopy to characterize solubilized proteins.

Production, purification and characterization of proteins from soluble fractions

Proteins obtained from soluble fractions are usually contaminated by other proteins from the expression host, and thus require purification. This involves chromatographic techniques in aqueous solutions, in particular affinity chromatography and gel filtration chromatography (both commonly used by ~80% of all participants in this study), as well as ion-exchange chromatography (commonly used by 70% of all participants) (Q18). Hydrophobic interaction chromatography and precipitation methods (*e.g.* using ammonium sulfate) are used to a much lesser extent (each is commonly used by <15% of all participants). Affinity chromatography usually involves short peptide sequences, or complete globular proteins, that are expressed as fusions with the protein that is to be produced by use of appropriate genetic vectors. The most abundant affinity tags are the His-tag and the GST (glutathione-S transferase) tag, which are commonly used by 81% and 30% of all participants, respectively. MBP (maltose-binding protein) and STREP tag, as well as other tags, are used to a much lesser extent (<10% of the participants commonly use these tags) (Q19). To confirm the identity of a novel purified protein, sodium dodecylsulfate polyacrylamide gel electrophoresis (SDS-PAGE) is performed by virtually all participants who answered this question (>80% of all participants). However, this method provides only a molecular weight estimate, and does not confirm the identity of a protein beyond doubt. More reliable are mass spectrometry techniques (routinely used by 60% of all participants), as well as Western blotting and protein sequencing (both routinely used by ~30% of all participants) (Q21).

Methods used to analyze the folded state of proteins in solution involve techniques similar to those used for analyzing the success of refolding processes involving proteins purified under denatured conditions to similar extents (Q20).

Methods used to improve the solubility and yield of expressed proteins

In order to increase the amount of soluble protein, various adjustments in the expression system, expression conditions and purification strategies have been described. The most common modification used by the groups participating in this survey is a reduction in the expression temperature (60%), followed by the use of solubilizing fusion tags and a reduction in expression levels (both 35%), codon optimization (29%), and a change in the expression system (23%). Less frequently employed to improve solubility were the purification from inclusion bodies in combination with refolding, coexpression of chaperones, the use of vectors with different promoters, and the exchange of the cellular localization (*e.g.* secretion into the periplasmic space) (Q22). As the relative use of the various strategies differs greatly, and a significant number of participants did not provide any information, it is difficult to evaluate their general usefulness in improving solubility. About 20% of the participants state that changing the expression system, codon optimization, reduction in expression levels, and the use of solubilizing tags have all successfully increased the amount of soluble material. Even more success is reported for the expression of proteins at decreased temperature, for which 40% of the participants stated frequent improvement in obtaining soluble material.

Further ways commonly used by the participating groups to improve solubility include optimization of buffer conditions, *e.g.* changes in salt concentration and/or pH and/or

the addition of additives (55%), expression of constructs based on homology to known domain structures (50%) or upon disorder and secondary structure predictions (36%), as well as the introduction of interaction partners (43%). To a lesser extent, constructs produced by limited proteolysis or homologous proteins are expressed to deal with solubility problems (commonly used by <20% of the participants). Only a very few groups try to improve solubility by systematic cloning of fragments of the full-length sequence or by the optimization of constructs making use of evolutionary methods (both <3%).

Labeling of proteins for X-ray crystallography

Solving the crystallographic phase problem requires methods such as MAD, SAD, MIR, and others. For these kinds of experiments the protein used for crystal growth needs to be labeled using heavy atoms such as heavy metals, xenon or halides. The most common methods still appear to be selenomethionine incorporation and heavy-atom soaking, which are frequently used by 36% and 26% of the participating groups, respectively (Q23). In contrast, xenon (Xe) labeling and halide soaking are frequently used by <6%, which may be due to the fact that these methods are relatively new and require special instrumentation, *e.g.* pressure cells for Xe. Selenomethionine is introduced into the protein almost exclusively by over-expression in *E. coli* (53 groups).

Labeling of proteins for NMR spectroscopy

Introduction of isotopes with suitable nuclear spins into proteins is required for the detection for changes in the chemical environment of certain groups, and for the determination of distances between nuclei and of their orientations relative to each other. Labels frequently used are ^{1}H (already present in all amino acids), ^{13}C, ^{15}N, and ^{31}P. For the introduction of labeled amino acids the protein is typically produced in *E. coli* (as reported by 25 groups) (Q23). Yeast strains (4 groups) and insect cells (1 group) have been used only rarely.

Use of personnel, collaborators, internal and external resources

Prior to protein production, the respective expression systems must be generated by cloning. For this study we distinguished three categories:

A. plasmid or PCR-product-based expression systems for protein production (this includes expression in bacteria, yeast, and cell-free expression systems).

B. transient eukaryotic expression systems (e.g. lipofection, as well as other methods of transfection and virus-based expression).

C. stable eukaryotic expression systems (eukaryotic cell lines featuring genome-integrated transgenes).

These three methods were used frequently by the participants in this survey (A used by 52, B used by 23, C used by 17 groups). Independent of the approach, cloning and expression experiments were usually carried out by group members, and to a significantly lesser extent by collaborators. In contrast, protein production by central facilities

within the institutions, by external companies or by non-profit production centers was reported to a much smaller extent: Less than 10% of all participants use such services, and if they do, only for selected projects (Q24).

The distribution of work is similar when it comes to protein purification (Q25). With respect to quality assessment (Q26), the use of central facilities within the institute is a little higher, but is commonly resorted to by <10% of the participating groups.

Rate-limiting steps in protein production and approaches to overcoming them

With respects to bottlenecks in protein production, the most limiting factor in the work flow appears to be the optimization of soluble protein expression (commonly mentioned by ~56% of the participants). This was followed by optimization of purification and optimization of refolding (both ~30%), cloning (16%) and quality assessment (11%) (Q27).

When asked whether specialized 'hands-on' courses for principal investigators and/or their coworkers would improve their productivity and success in protein production, ~53% of the participants in this study agreed that this would be helpful, and only ~10% did not feel that this would be beneficial (Q28). The priorities for such courses are: optimization of expression, protein purification, protein characterization, protein refolding, cloning, and protein labeling. With the exception of the latter subject, at least 40% of all participants are convinced that their productivity would be improved by such courses, indicating that there is indeed a major demand for additional training and support in these areas. We therefore suggest that specialized courses open to and organized by structural biologists from a wide range of backgrounds could increase the speed and efficiency of structural analysis by the exchange of knowledge, providing information about new developments, and providing practical experience.

Another measure of support for groups working in SG/SP and SB would be the establishment of a database on protein production, as critical information which has been acquired is frequently not available to the community as a whole. This lack of data results in redundant work by more than one research group. Based on suggestions from the participants, such an initiative should make available information complementary to that provided by standard protein databases, namely information on expression systems, purification protocols, literature references, yields of protein production experiments, information concerning failed expression trials, links to groups working on expression and purification of particular proteins, information concerning biological activities of proteins and protein-ligand interactions, and access to relevant biological materials. All these types of information are considered to be essential or important by at least 40% of the participants in the study (Q29). Provision of raw data such as gel images and blots was considered to be essential or important by only ~23% of the participants. If available, such a database would be used by ~70% of all participants, and only ~6% expected that it would not play an important role. On average, the participants who contributed information stated that they would be willing to devote 30-45 minutes to entering data for a particular protein into such a database.

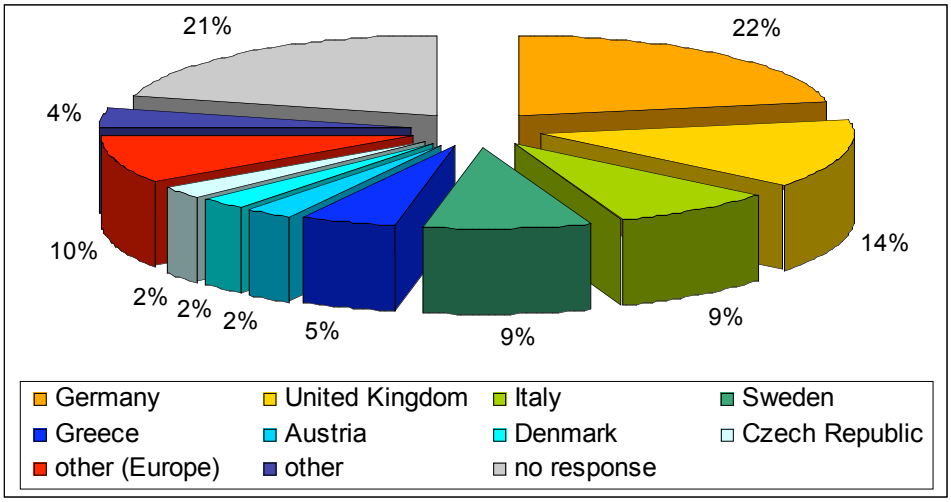
The development of a standardized vector platform and outsourcing of laborious repetitive tasks to non-profit service centers also promises to greatly relieve structural biologists from the burden of time-consuming routine work. 61% of the participants in

the study would make use of such services, if they were available, whereas 17% stated they would not (Q30). A vector platform tailored for SG/SP work would permit fast cloning and recloning of genes into expression vectors, make available a repertoire of fusion tags, permit overexpression of genes in various organisms, and overcome limitations of some commercial systems, such as additional amino acids resulting from recombination sites. Services provided by these non-profit service centers might involve design and/or evaluation of cloning strategies, generation of custom-made genetic constructs, generation of protein expression systems, and the testing and optimization of protein production. All these services were considered to be essential and important by about half of the participants (Q30).

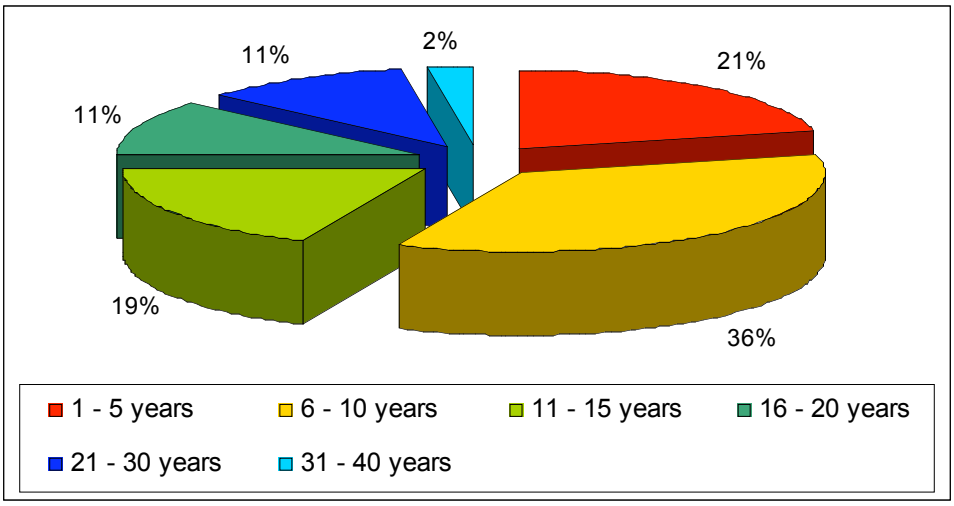
Based on the responses to this survey, systems for recombinant expression provided by non-profit service centers should include plasmid-based expression in *E. coli* (51%), plasmid-based expression in yeast (39%), and baculovirus-based expression in insect cells (40%). Surprisingly, the production of eukaryotic cell lines stably overexpressing genes of interest was considered less important (21%). This may be due to the fact that use of eukaryotic cell lines is still rather new to structural biologists, and their usefulness might be judged more positively as they become more widespread. Services concerned with protein production should be involved in testing and optimization of protein expression (51%), protein purification (43%), and protein refolding (29%).

Survey Questions and Answers

Q1 Participating groups

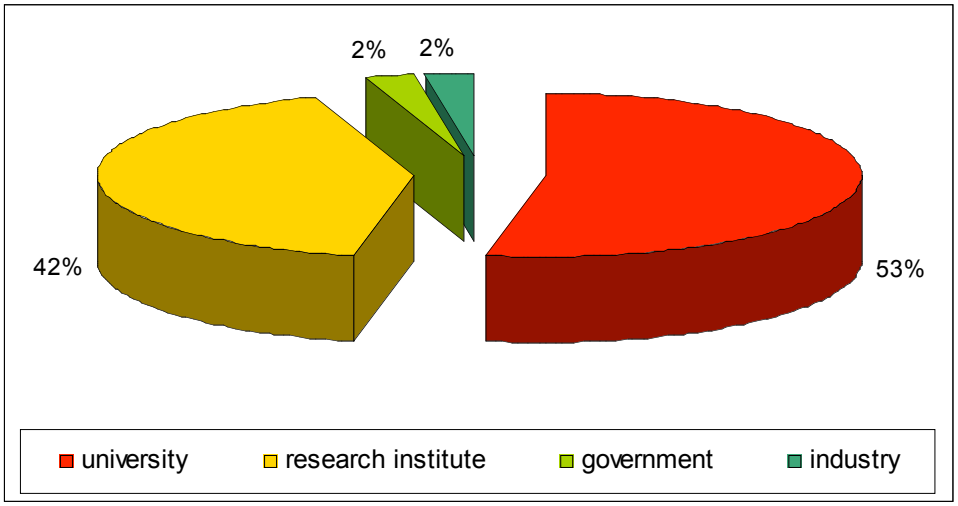


Q2 Years as primary investigator:

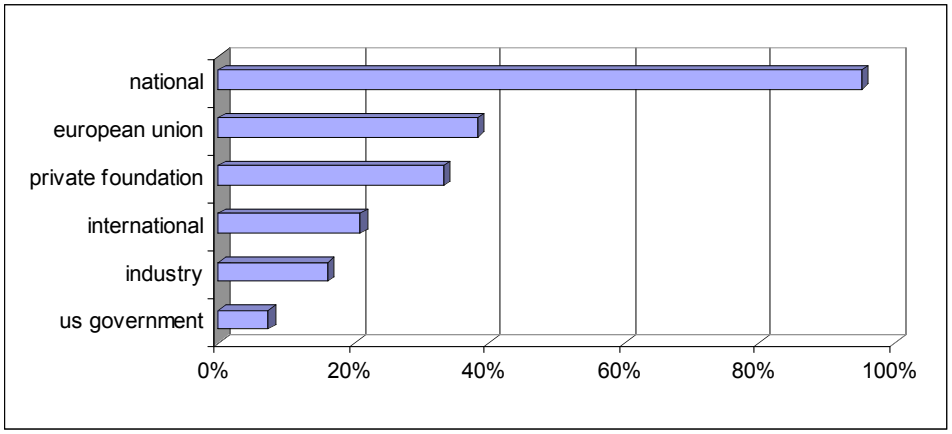


median: 10 years

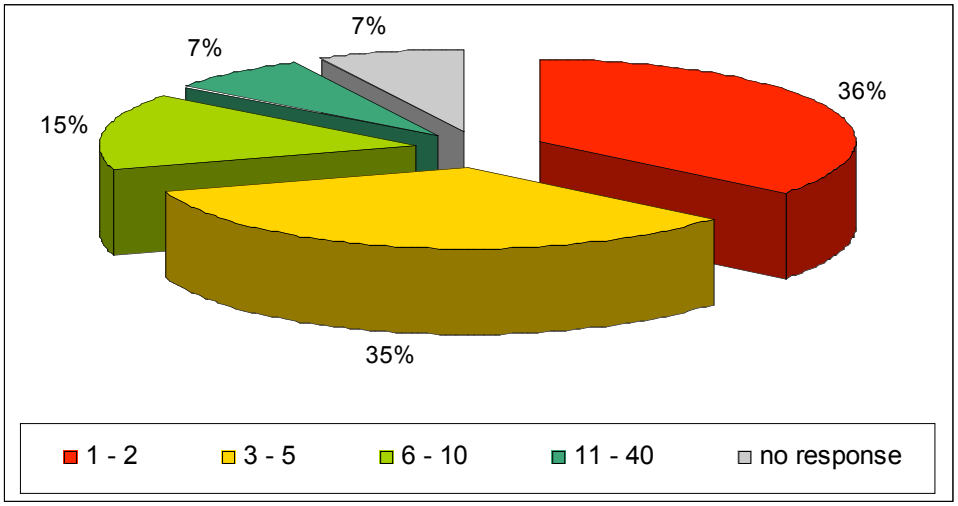
Q3 Research environment



Q4 Research support

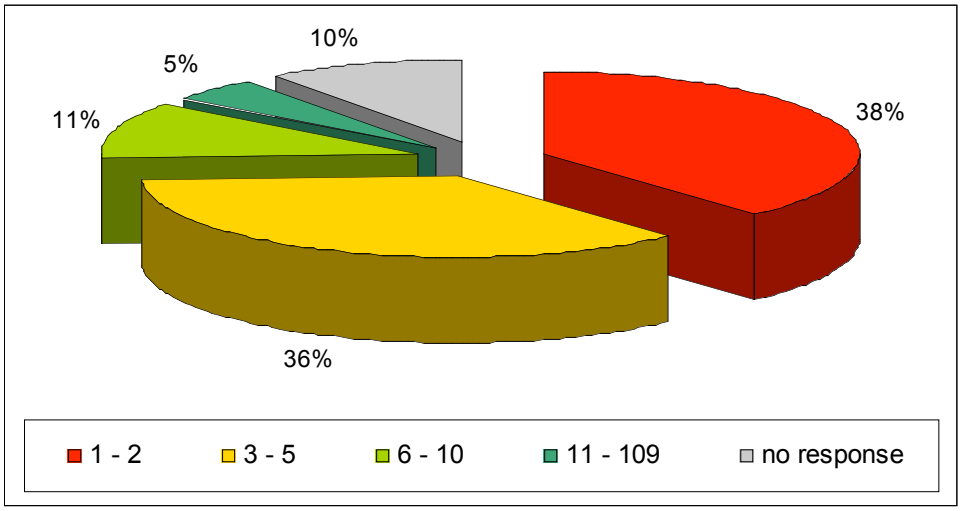


Q5 Number of postdocs & research associates



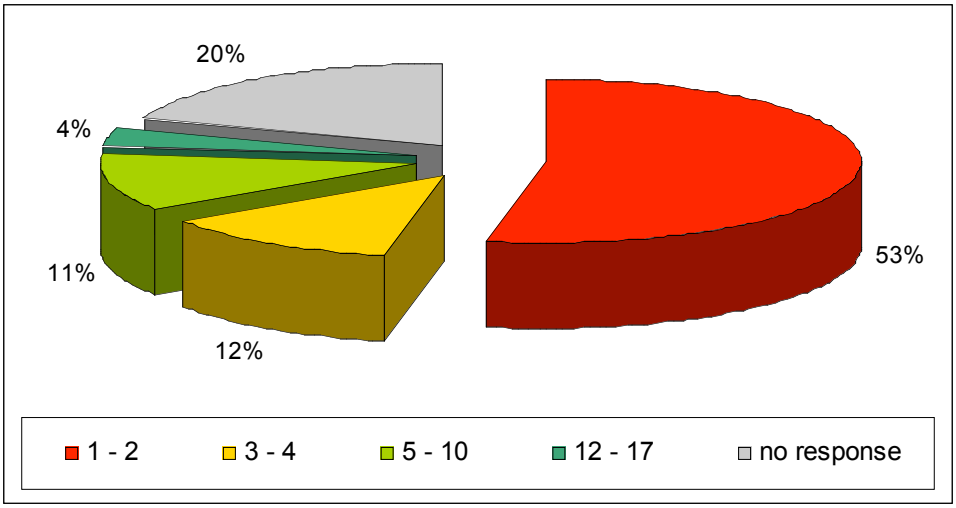
median: 3

Number of graduate students



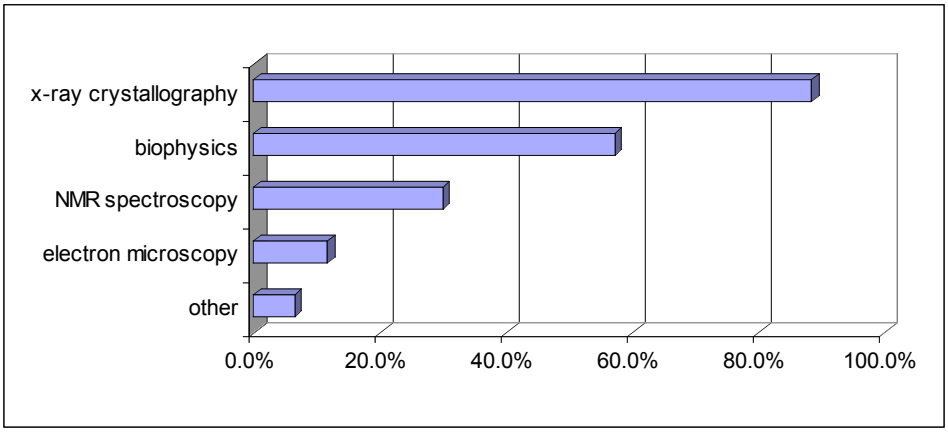
median: 3

Number of support staff:

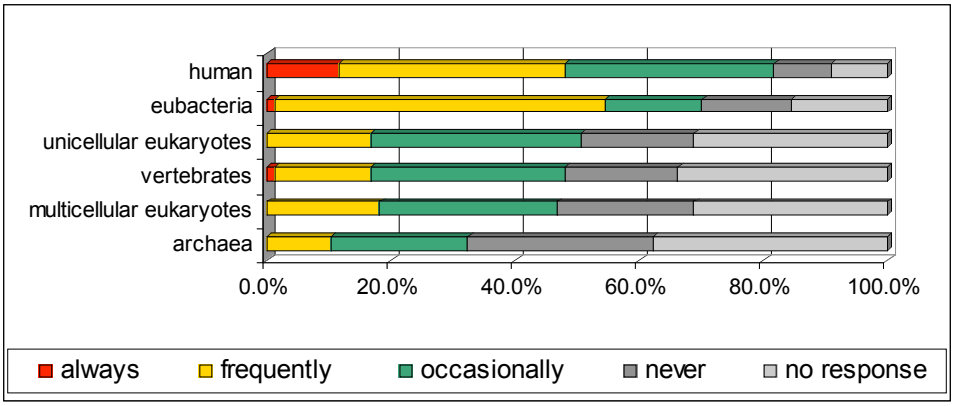


median: 2

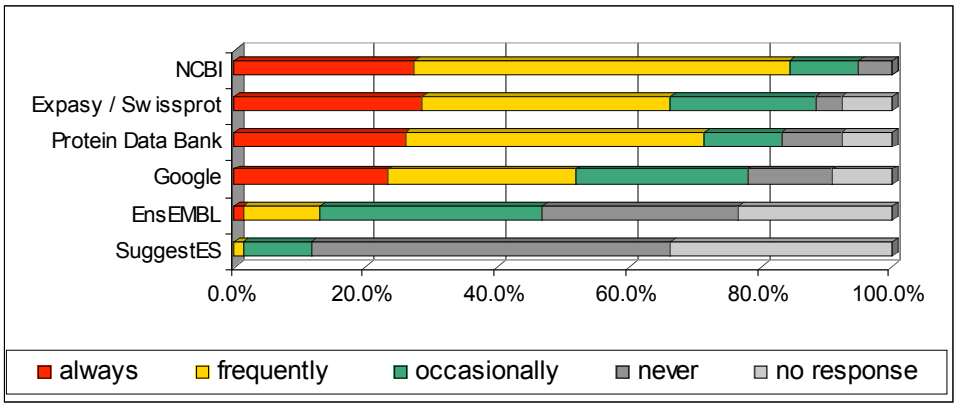
Q6 Which are the main experimental techniques used in your group?



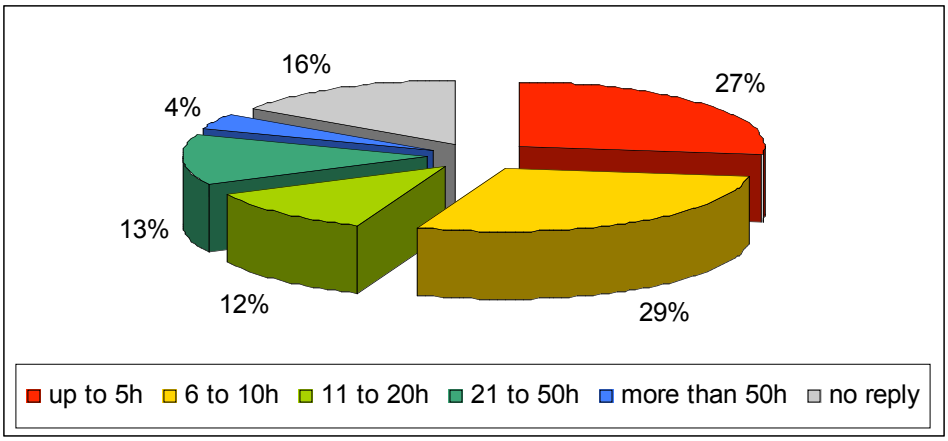
Q7 What are the source organisms of the proteins you study?



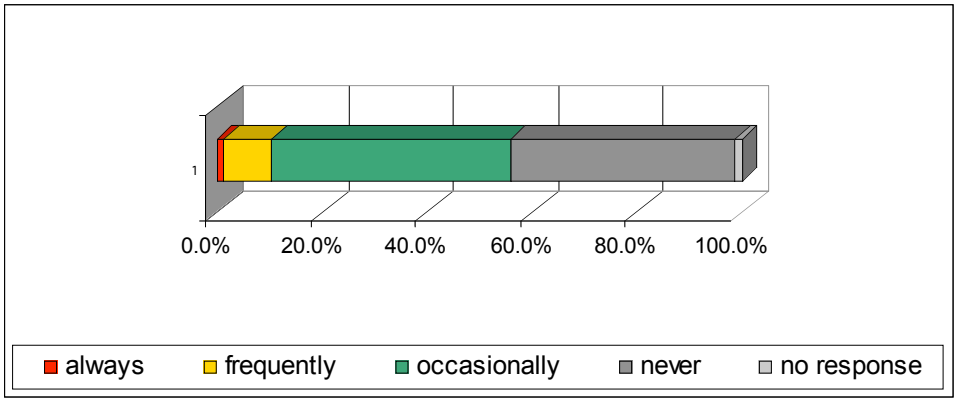
Q8 Which internet-based databases /search engines do you use to find information on protein cloning and expression ?



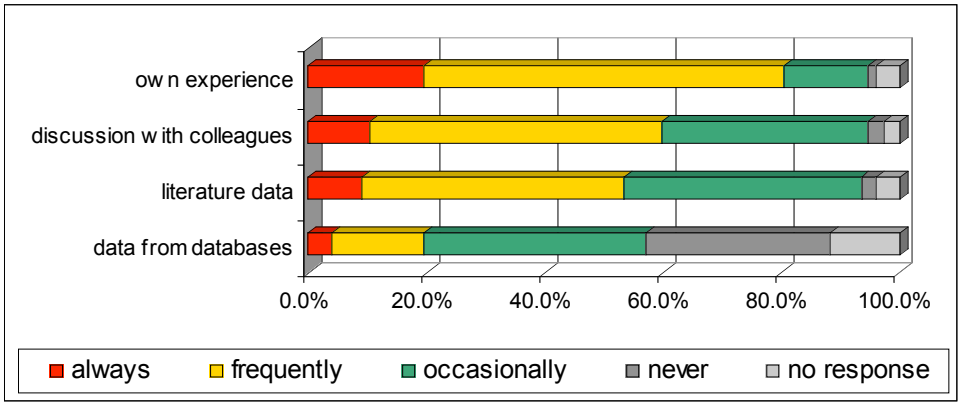
Q9 How much time do you invest on literature research before starting the production of a novel protein?



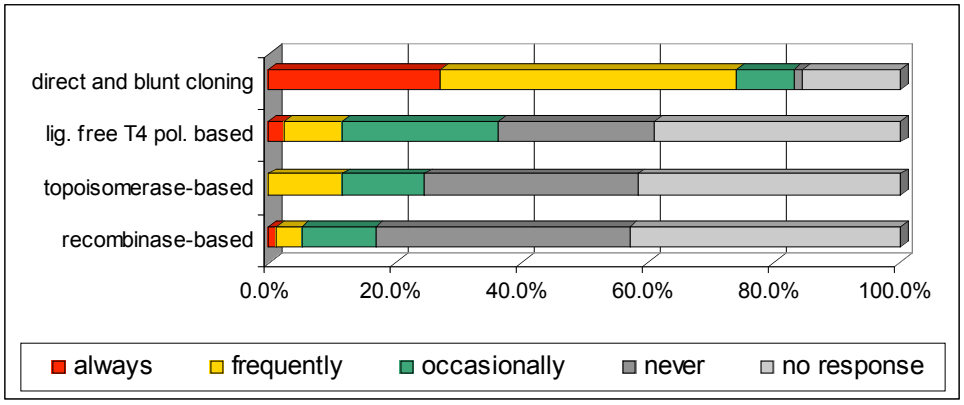
Q10 Do you use commercial providers for a protein / clone of interest?



Q11 How do you select an expression system for a protein?

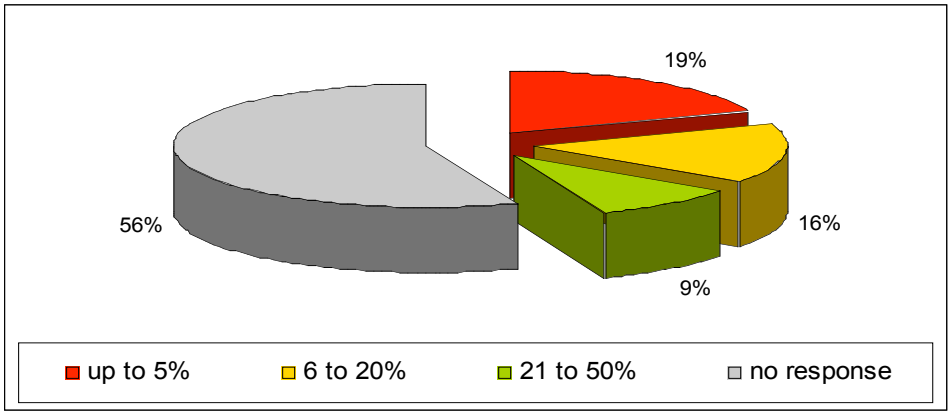


Q12 If applicable, which cloning methods do you use?

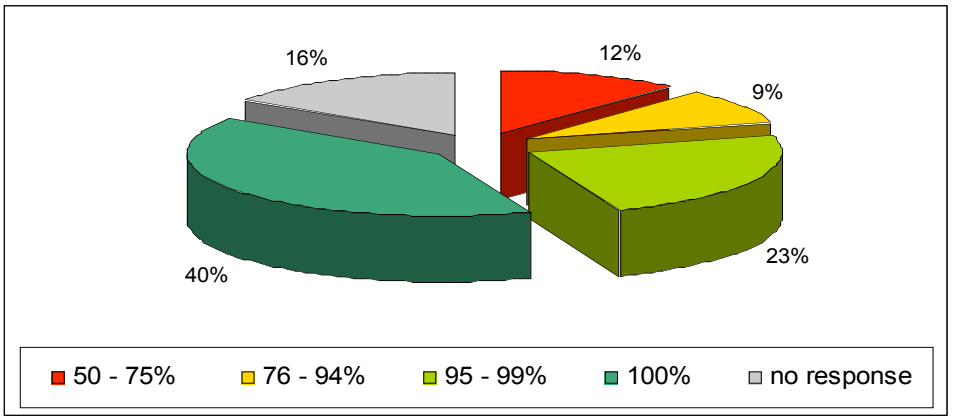


Q13 Which techniques do you use for protein production?

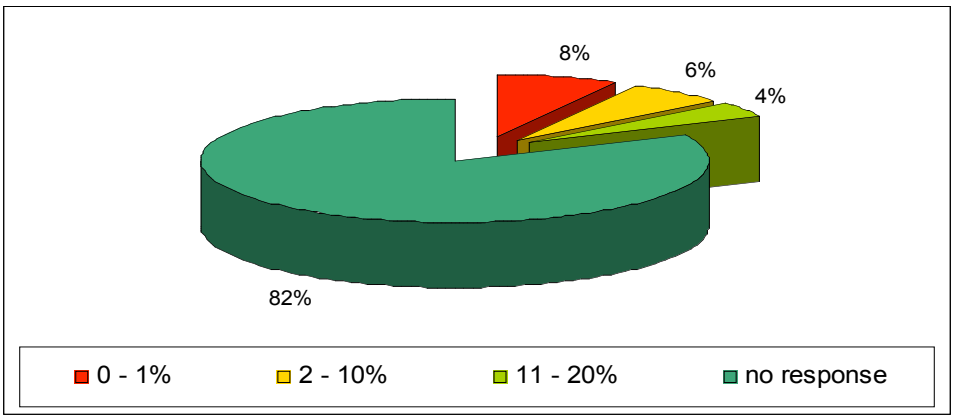
Isolation from native sources, total responses: 34, average use by participants: 14%



Recombinant expression, total responses: 65, average use by participants: 94%

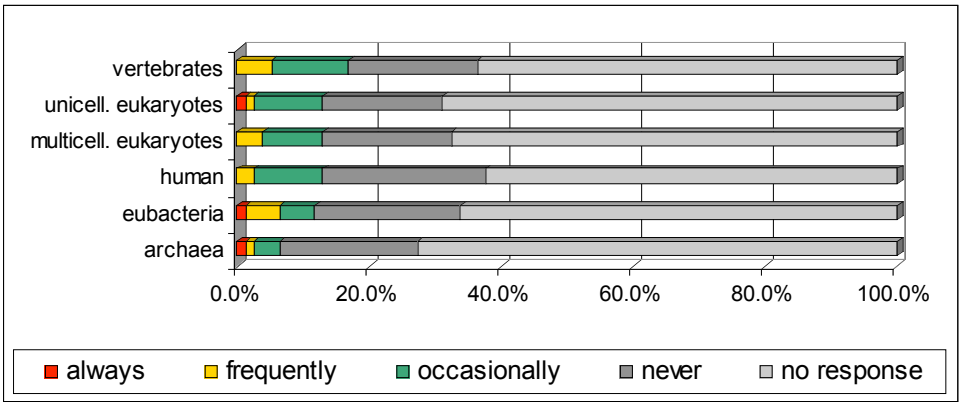


Cell-free expression, total responses: 14, average use by participants: 6.4%

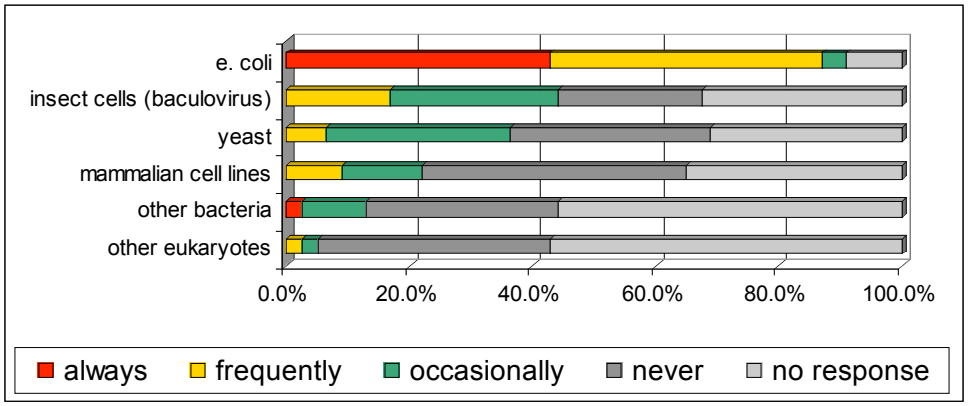


Q14 Which organisms / sources do you use for protein production?

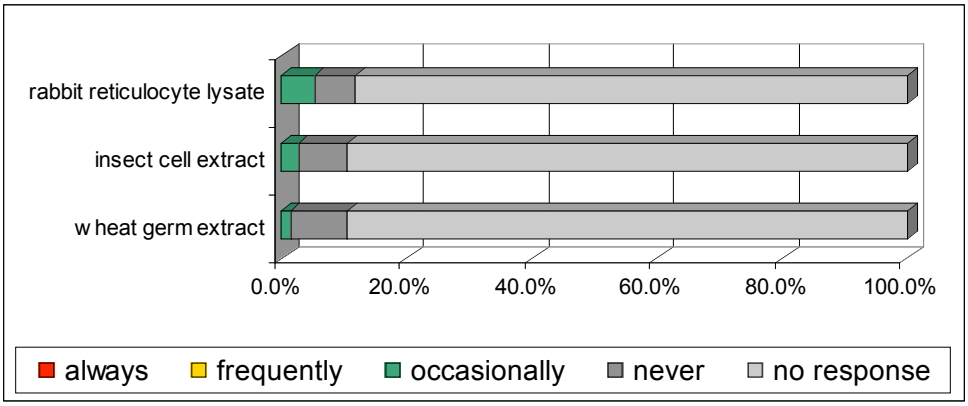
Isolation from native sources:



Recombinant expression:

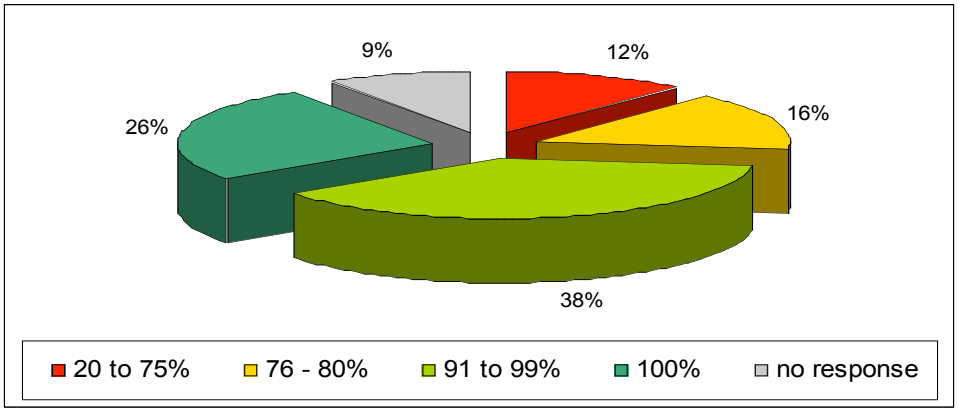


Cell-free expression :

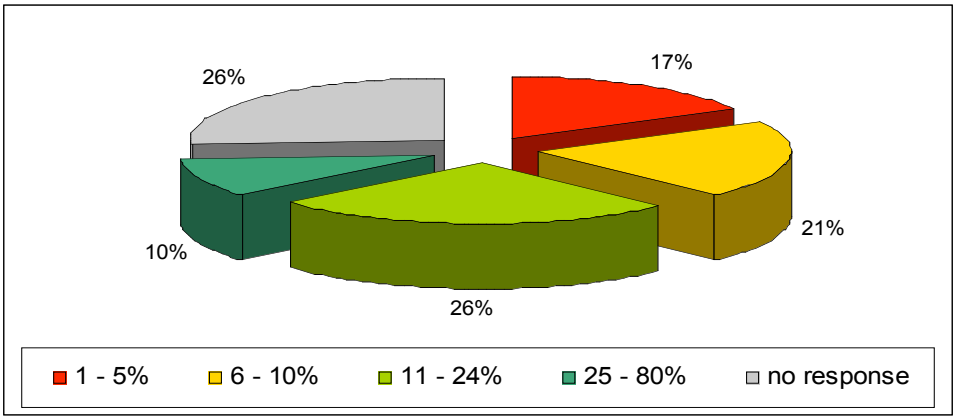


Q15 Protein purification in your group is performed under native or denaturing conditions ?

Purification under native conditions, total responses: 70, average use by respond.: 88%

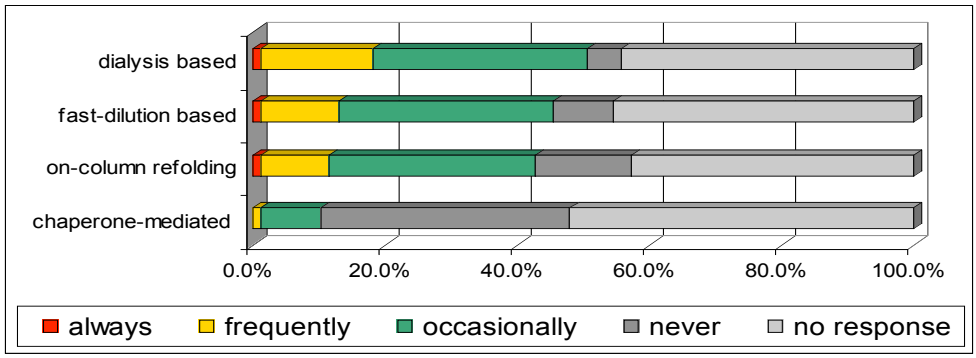


Purification under denat. conditions, total responses: 57, average use by respond.: 14%

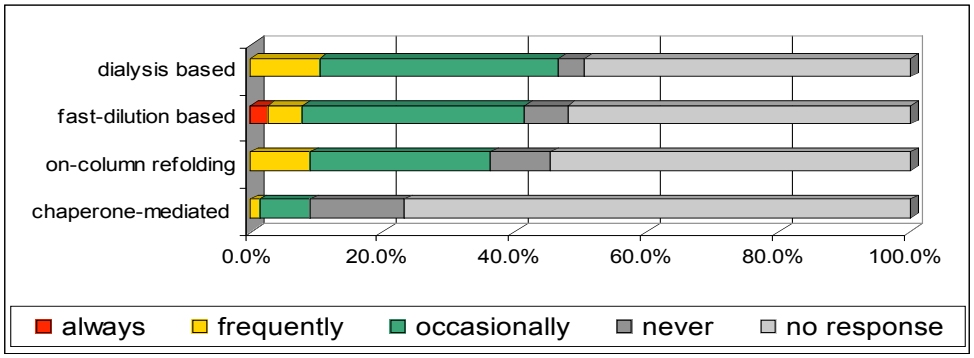


Q16 Which methods do you apply to refold proteins and how successful have they proven to be?

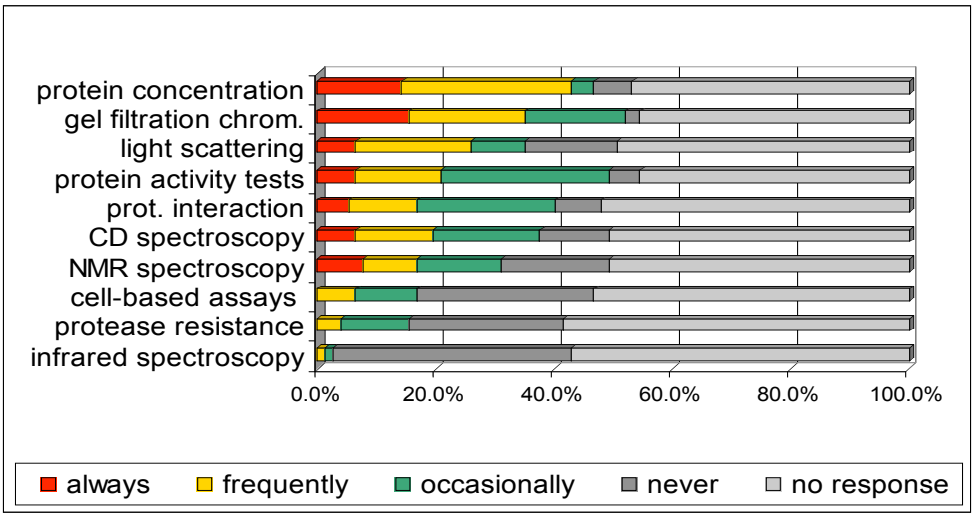
Use:



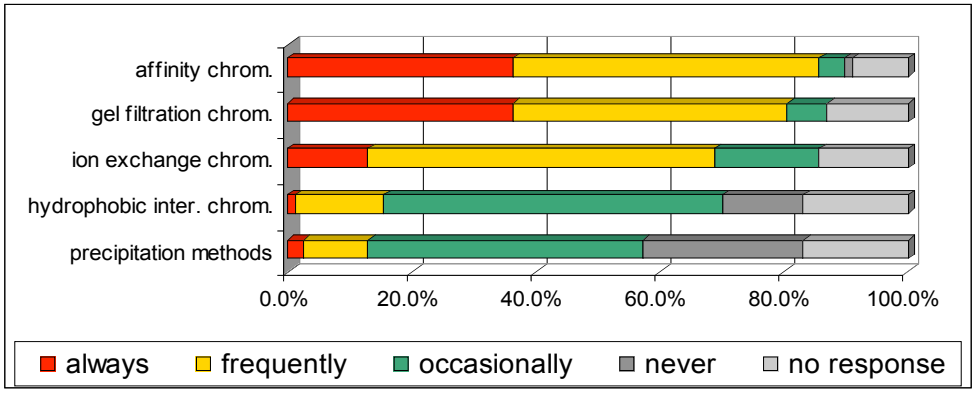
Success of refolding:



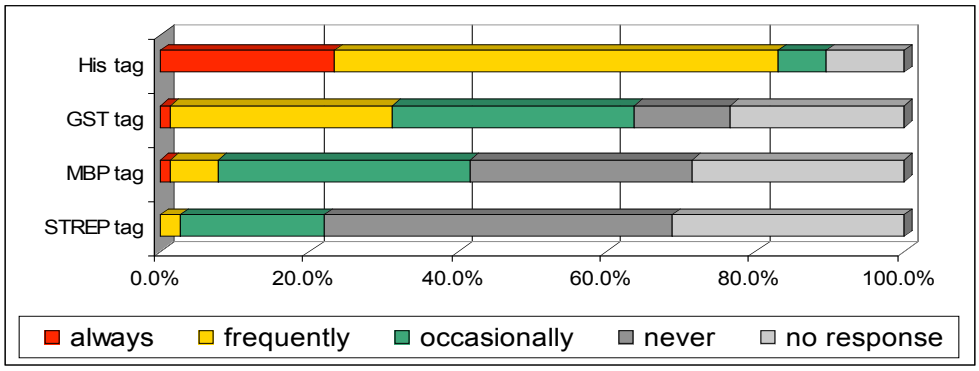
Q17 How do you monitor the success of refolding?



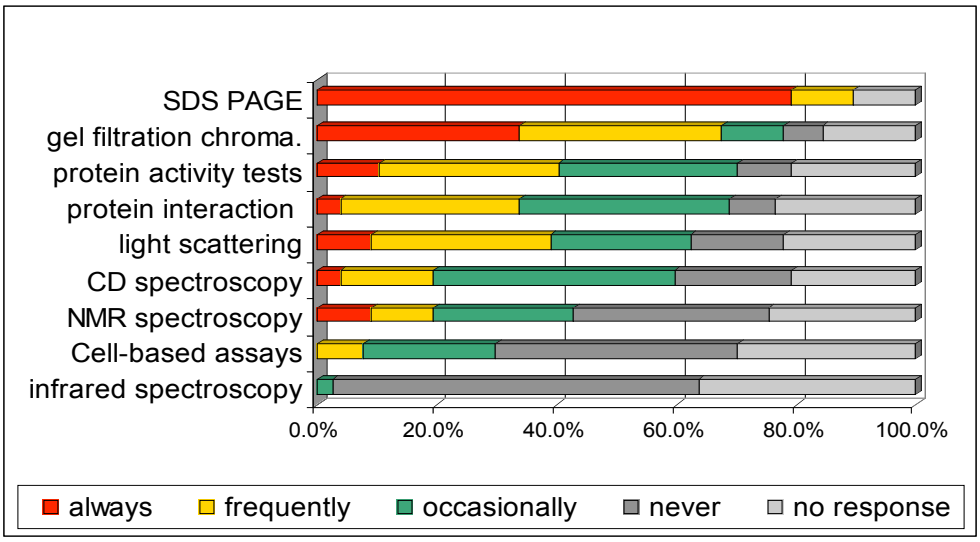
Q18 Protein purification in your lab routinely involves



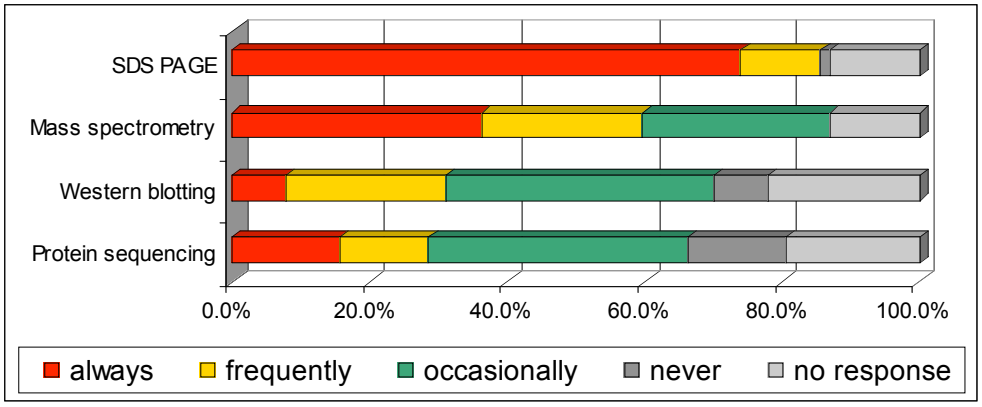
Q19 Which affinity tags do you use?



Q20 Quality assessment of a purified protein before experimental use

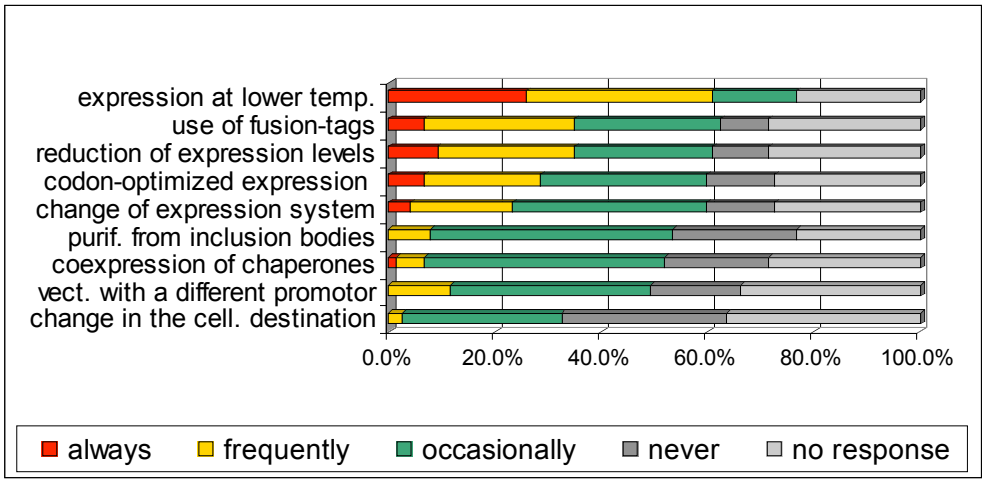


Q21 How do you confirm the identity of a novel purified protein?

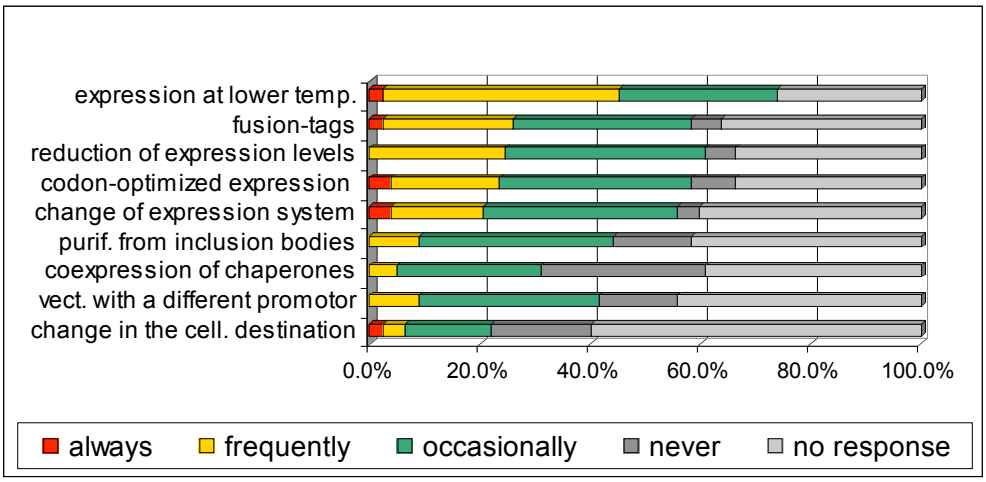


Q22 Which strategies do you use to improve protein solubility and / or prevent protein aggregation, and how successful have they been in your work?

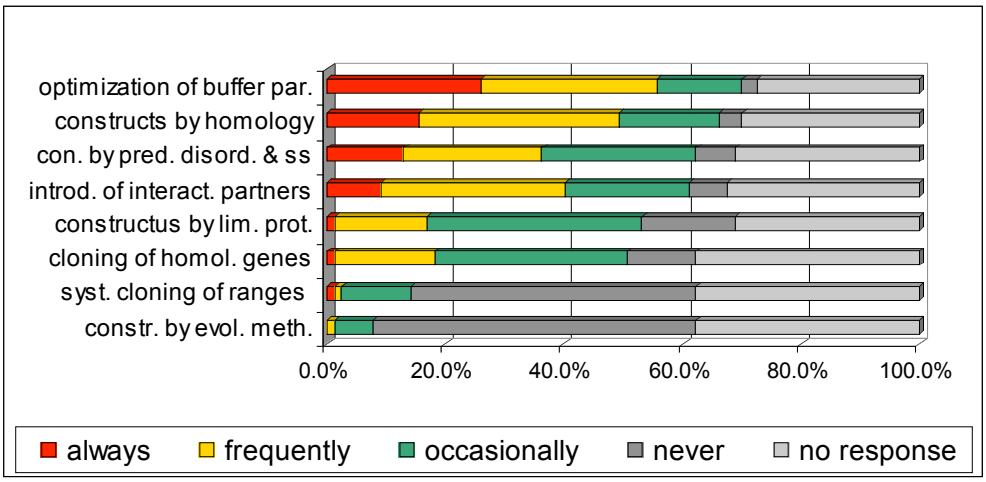
Use: A. Variation of expression conditions and expression systems



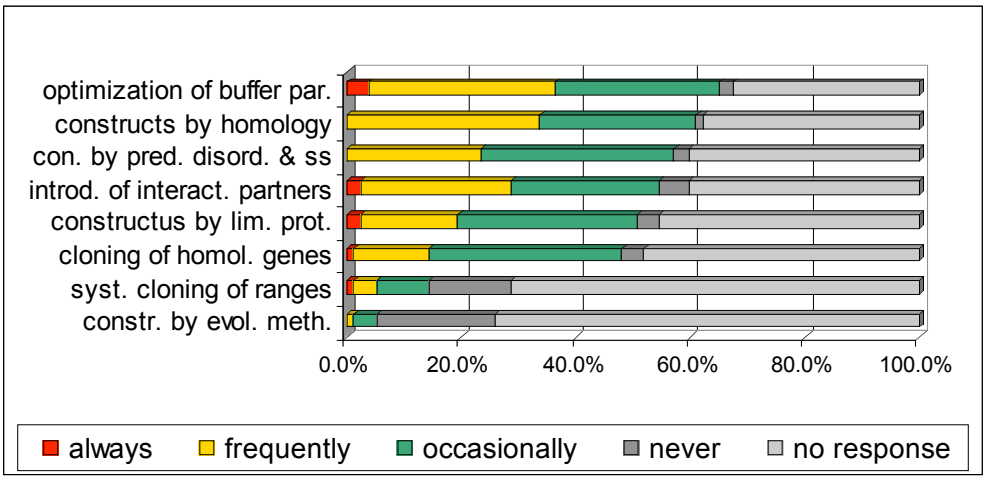
Success: A. Variation of expression conditions and expression systems



Use: B. Optimization of the molecular environment and construct design



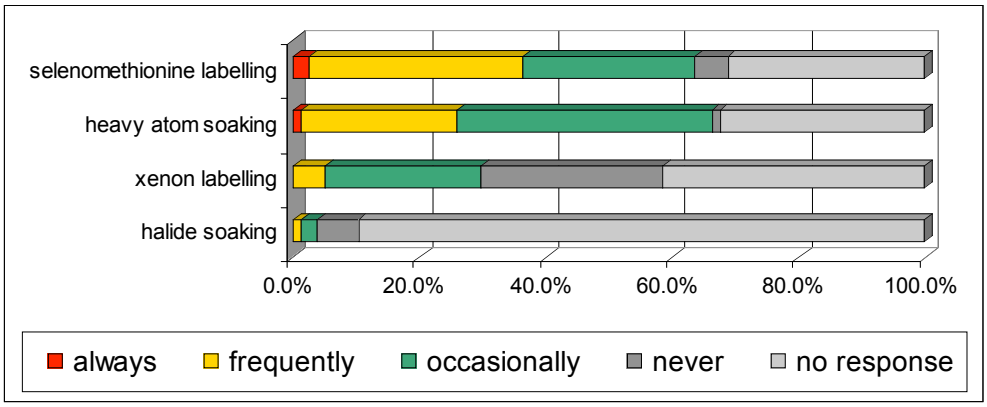
Success: B. Optimization of the molecular environment and construct design



Q23 Which labeling technique do you use in protein production?

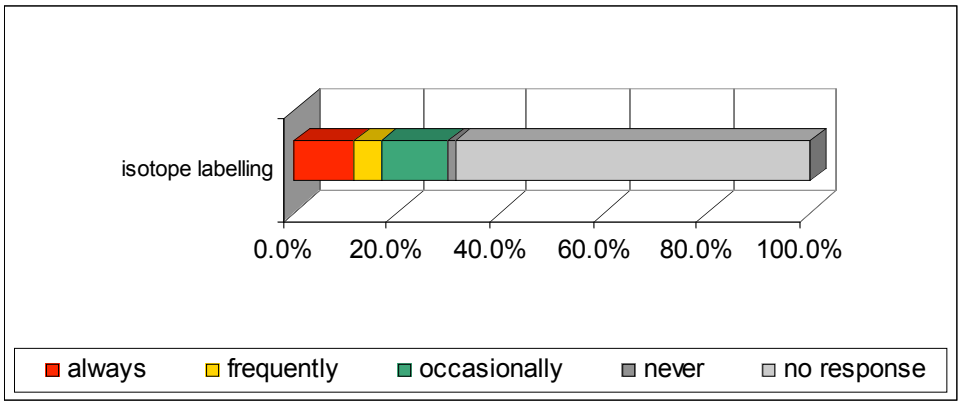
(yes, no, no response):

X-ray crystallography: applicable: 53, 8, 16



(yes, no, no response):

NMR spectroscopy: applicable: 25, 31, 21

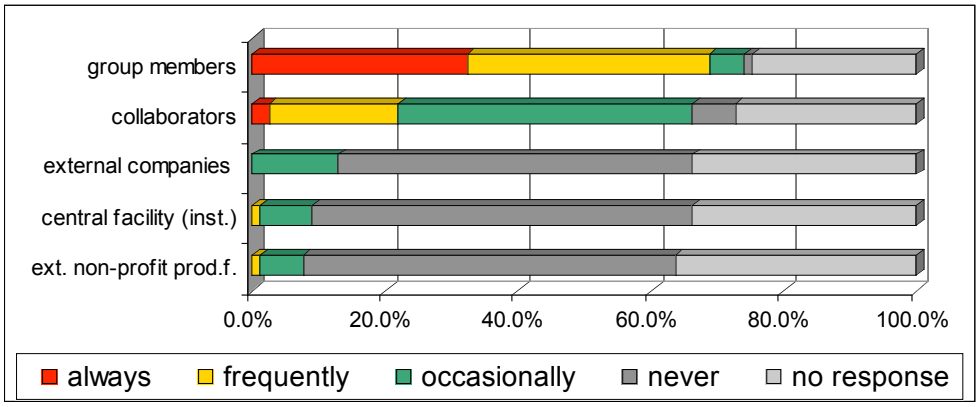


Q24 Cloning and establishment of expression systems

A. plasmid or PCR-product-based expression systems for protein production (expression in bacteria, yeast and cell-free expression)

yes, no, no response
In use: 52, 3, 23

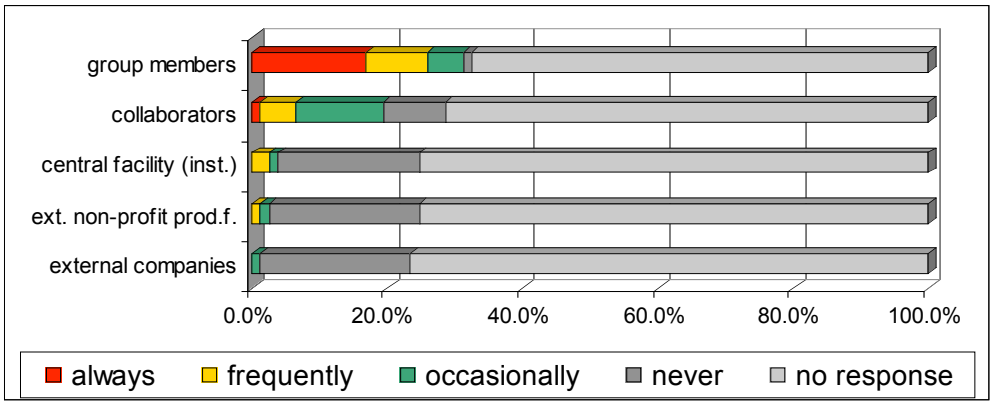
Work is performed by:



B. transient eukaryotic expression systems (lipofection and other methods of transfection and virus-based expression)

yes, no, no response
In use: 23, 30, 34

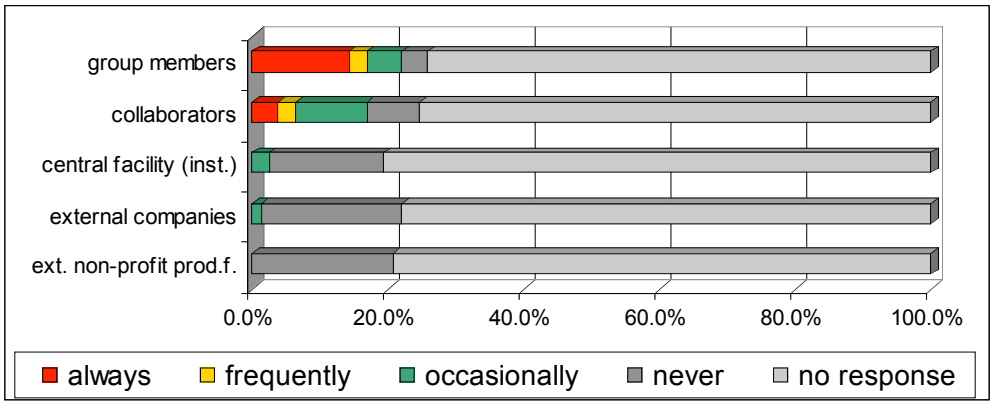
Work is performed by



C. stable eukaryotic expression systems *(eukaryotic cell lines featuring genome-integrated transgenes)*

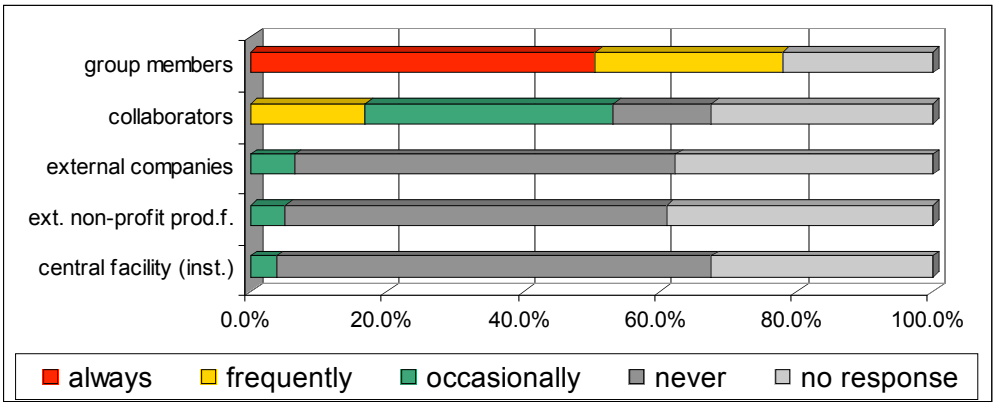
yes, no, no response
In use: 17, 33, 27

Work is performed by:



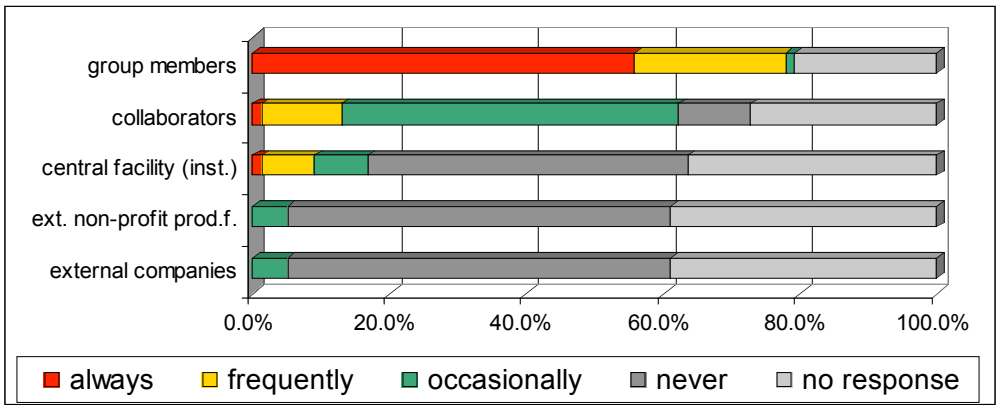
Q25 Protein purification

Work is performed by:

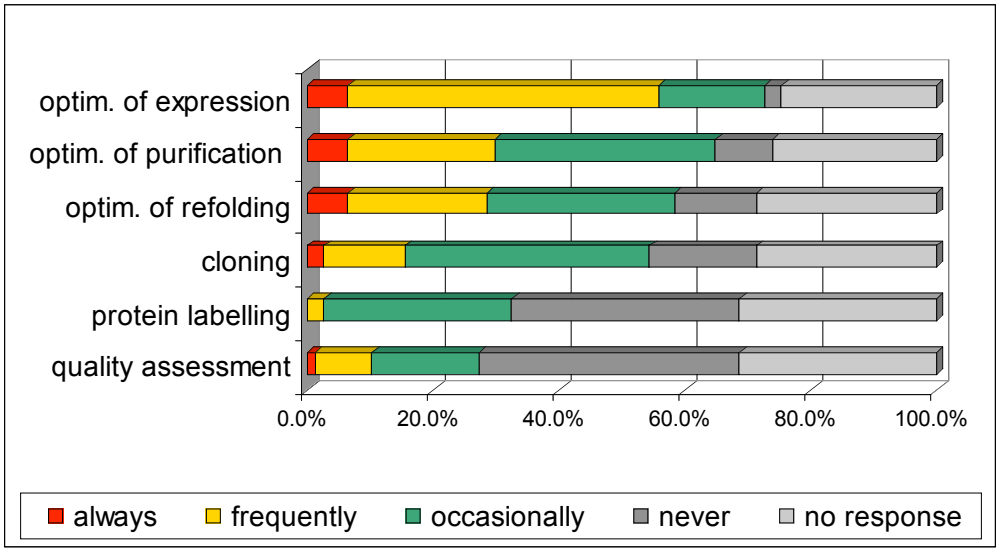


Q26 Protein quality assessment

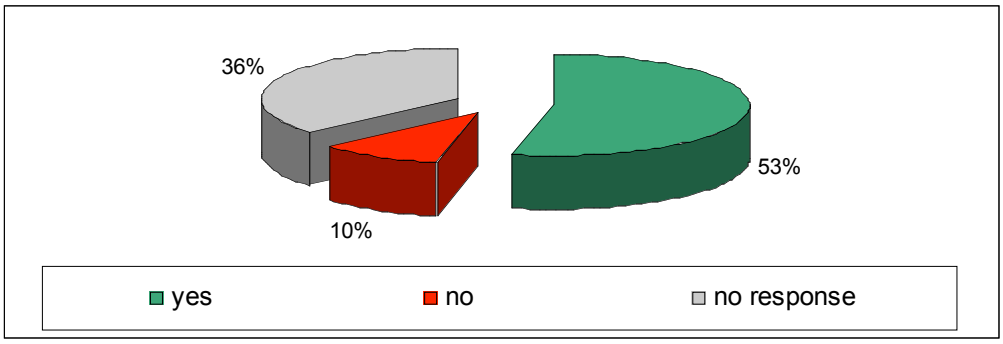
Work is performed by:



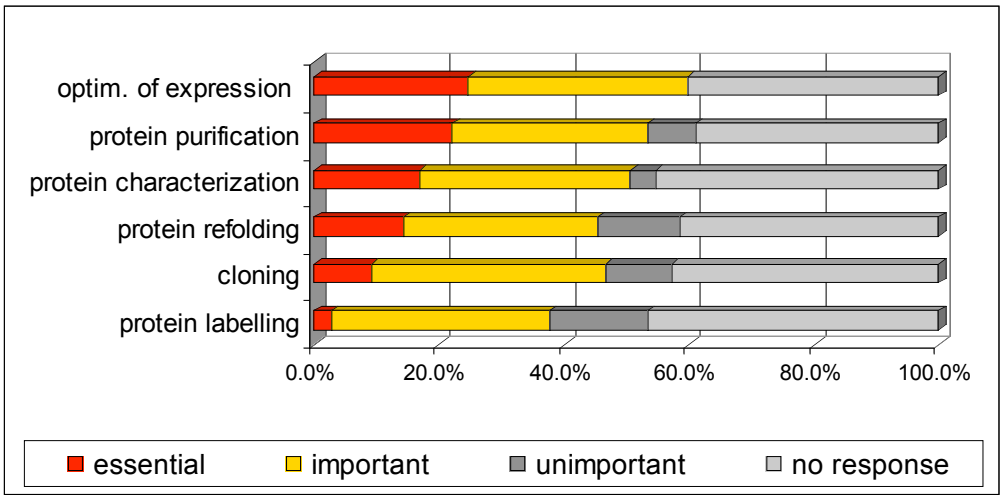
Q27 Which steps do you consider to be rate-limiting in protein production ?



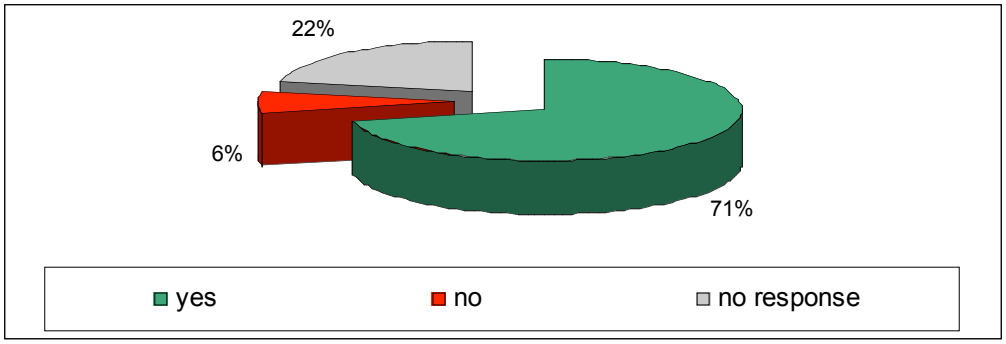
Q28 Do you think that specialized hands-on courses for you / your coworkers would improve your productivity / success in protein production?



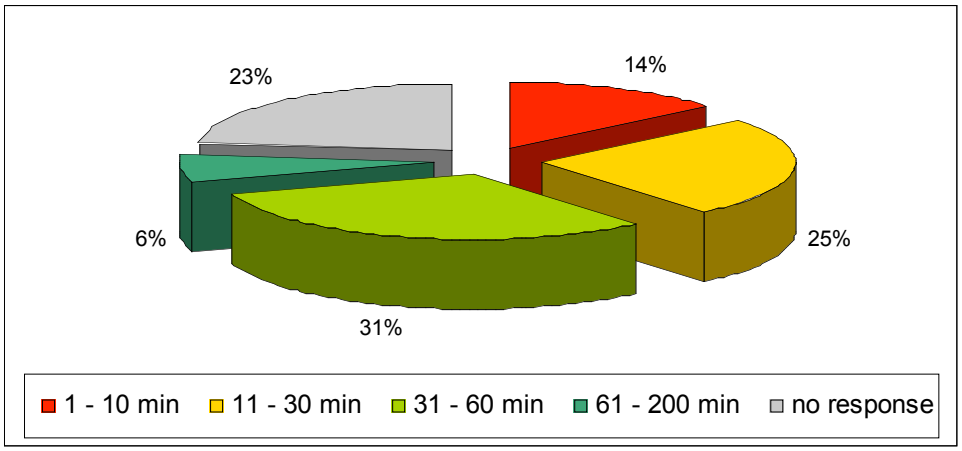
Which subjects would you consider of greatest relevance for such a training?



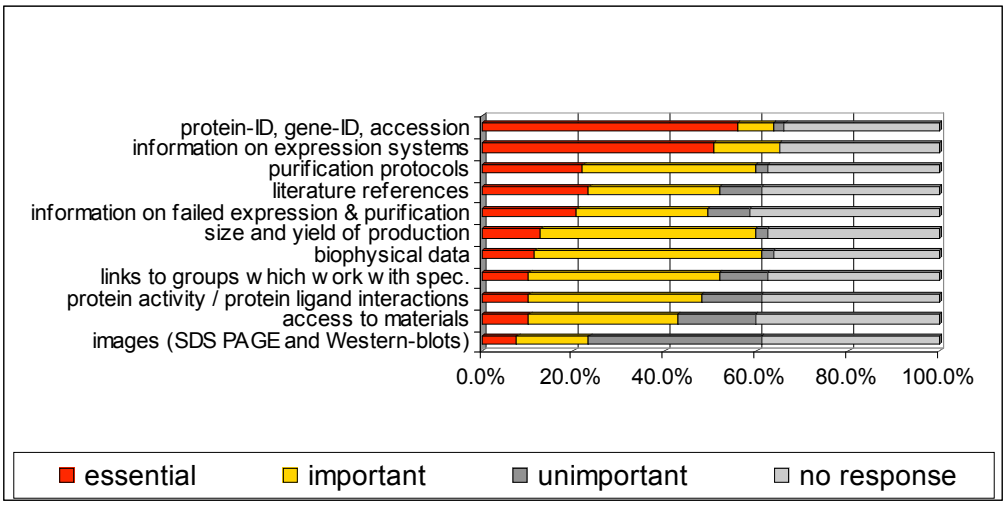
Q29 Would you use a specialized database on protein production, if available?



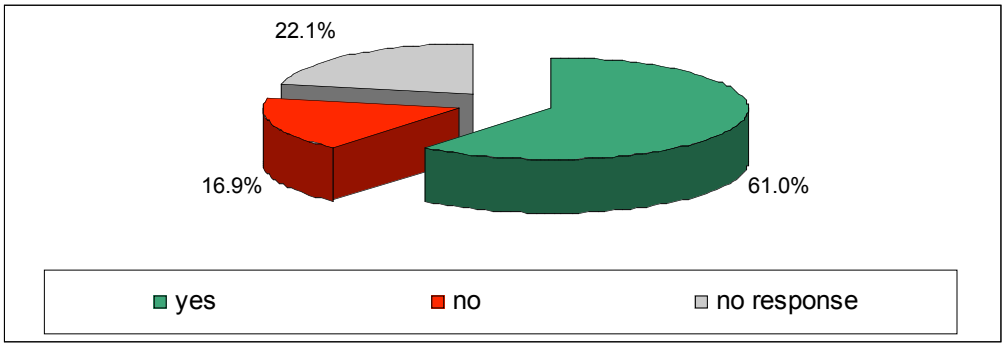
If yes, how much time would you be willing to invest for data deposition per protein?



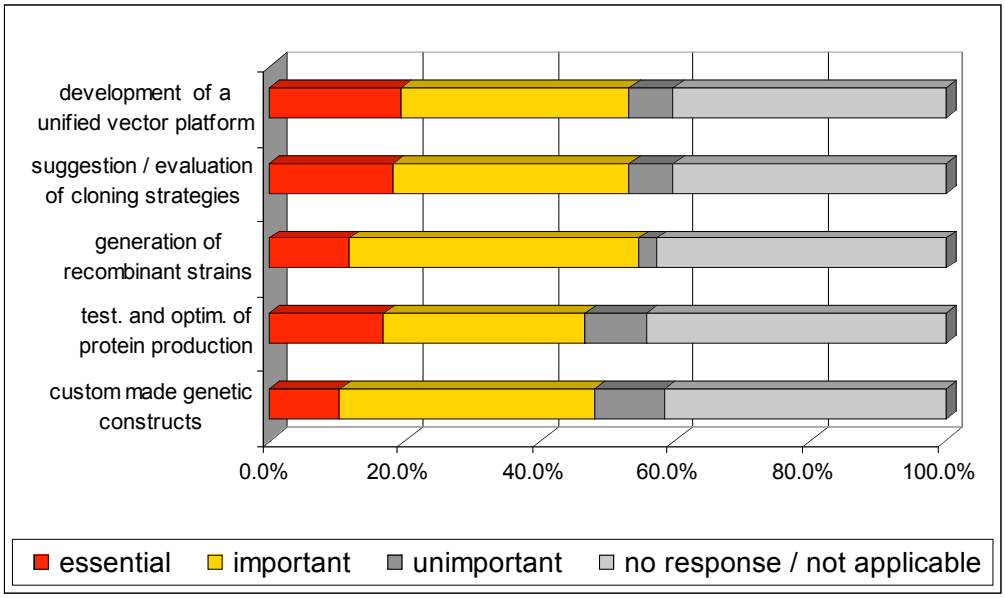
If yes, which information should be included / would be provided by you from your work for public access?



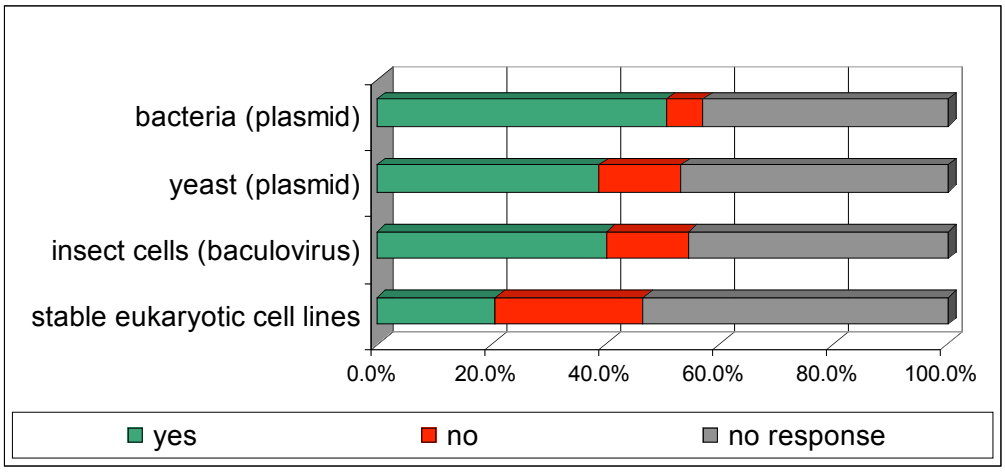
Q30 Would you use the service of specialized non-profit centers for gene-cloning / protein production, if available?



Which tasks / services should be provided by these centers?



Which recombinant strains for protein production should be generated by non-profit service centers ?



Testing and optimization of protein production should involve