Journal of
Applied
Crystallography
ISSN 0021-8898

Likelihood of crystallization: experimental and computational approaches

A report on the VIZIER Workshop on the 'Definition of protein domains and their likelihood of crystallization', Vienna, Austria, 28–30 June 2006.

1. Introduction

The principal reason to organize a workshop on likelihood of protein crystallization was to bring together various competencies, ranging from crystallography and NMR spectroscopy to high-throughput protein production and bioinformatics.

Such a diverse set of expertise has evident difficulty in interfacing because of the different scientific procedures and approaches that are followed in these scientific fields. For instance, new bioinformatics methods are validated on existing experimental data, which are deposited in various databases, while novel experimental procedures are in general validated with respect to their ability to produce new data, not yet included in databases. The likelihood of crystallization is thus predicted, *via* bioinformatics approaches, on the basis of known facts, while it is experienced, by structural and molecular biologists, on the basis of new observations.

The main topics of the workshop – conformational disorder, protein domain boundaries and post-translational modifications, and associated problems, which may hinder a successful structural determination – were dissected from both the experimental and the computational perspective.

The workshop was started by a lecture of Dmitrij Frishman (Technical University, Munich, Germany), who summarized the recent achievements of his laboratory in designing methods for predicting the crystallizability of proteins on the basis of their sequences. This lecture was, in part, a general introduction to the workshop, given that several bioinformatics applications are devoted to the general problem of predicting if a certain sequence will be suitable for expression, purification and crystallization. Machine learning methods were used to distinguish proteins that were crystallized from proteins that were experimentally proven to be difficult to express or crystallize, and the prediction accuracy reached values of about 60–70%, depending on the learning and test sets. It was agreed that such performance is still far from satisfactory for practical purposes, though the computational strategy seems to be very promising and the prediction quality is likely to improve significantly when larger data sets become available. In particular the approach may be valuable in structural genomic projects when there is a desire to rank targets according to likely success. Other software suites and collections of Web-based servers that may facilitate structural biology experiments were presented by Jaime Prilusky (Weizmann Institute of Science, Israel), Oxana Galzitskaya (Russian Academy of Sciences, Pushchino, Russia) and Christoph Meier (Oxford Protein Production Facility, UK). Structural biologists can access these services to handle their specific problems and optimize the experimental strategy.

2. Conformational disorder

It is obvious that the three-dimensional structure of a conformationally disordered protein cannot be determined, either experimentally or computationally, and several talks were therefore focused on the prediction of conformational disorder on the basis of the amino acid sequence. Anne Poupon (Université Paris-Sud, Paris, France) described the PRELINK algorithm, which is based on amino acid composition and on hydrophobic cluster analysis. Zsuzsanna Dosztányi (Hungarian Academy of Sciences, Budapest, Hungary) summarized the IUPred method that predicts protein disorder by estimating the pairwise energy content from the amino acid sequence. A related method was presented by Robert Konrat (University of Vienna, Austria), who also discussed prediction with experimental NMR observations. A neural-network-based technique, RONN, was summarized by Christoph Meier (Oxford Protein Production Facility, UK). Oxana Galzitskaya (Russian Academy of Sciences, Pushchino, Russia) presented a technique based on predicted packing density. David Jones (University College London, UK) described the DISOPRED2 predictor, based on machine learning methods, which seems to be particularly suited for predictions of long disordered polypeptide regions. A detailed comparative overview of the various prediction methods was presented by Sonia Longhi (CNRS et Universités Aix-Marseille, France), who pointed out that, since none of the available methods for disorder prediction can be taken as fully reliable on its own, it is necessary to consider their advantages and drawbacks and to combine them to avoid pitfalls and to achieve more reliable prediction. Zsuzsanna Dosztányi (Hungarian Academy of Sciences, Budapest, Hungary) also reviewed the most recent experimental techniques based on two-dimensional gels, which allow one to identify on a proteomic scale conformationally disordered proteins

3. Domain boundaries

Another problematic issue in practical structural biology is the selection and fine tuning of the amino acid construct that is amenable to experimental analysis. It is in general a major concern for large proteins, usually composed of several separate structural domains, where the identification of the domain boundaries is often a crucial step along the entire experimental pipeline. Following the general workshop philosophy, both computational and experimental approaches were presented by the speakers. On the bioinformatics side, Alexander (Sasha) Gorbalenya (Leiden University Medical Center, The Netherlands) described how comparative sequence analysis can be exploited successfully for dissecting polyproteins of RNA viruses. Oxana Galzitskaya (Russian Academy of Sciences, Pushchino, Russia) presented a technique (DomPred) based on propensities of amino acid residues for domain boundaries. Amino acid propensities were also used in a method described by Christine Elsik (Texas AM University, USA) that uses hidden Markov models. A residue propensity scale for being in linker regions, intercalated between structural domains, was also used in the Armadillo procedure presented by Michel Dumontier (Carleton University, Ottawa,

Canada). Jaap Heringa (VrijeUniversiteit, Amsterdam, The Netherlands) reviewed a series of techniques based on consistency of multiple tertiary structure predictions (SnapDRAGON), on sequence homology (Domaination), on sequence hydrophobicity patterns (SCOOBY-DOmain) and on multiple sequence alignments (Sequence Harmony). Another two techniques, DomSSEA, based on the alignment of secondary structural elements, and DPS, based on amino acid sequence alignments, were described by David Jones (University College London, UK). Jaime Prilusky (Weizmann Institute of Science, Israel) presented a method (FoldIndex) for estimating the probability that a protein sequence is intrinsically unfolded.

It appeared that none of these computational approaches can offer, for the moment, highly accurate solutions that can be used on a routine basis to identify structural domains on the basis of protein sequences. In particular, the difficult problem of proteins containing more than two domains and of domains constituted by more than a single polypeptide chain seems to be far from solved. For this reason, experimental approaches are absolutely necessary to complement and confirm computational predictions.

Arnaud Poterszman (Universite Louis Pasteur, Strasbourg, France) showed several applications of limited proteolysis on the transcription/DNA repair factor TFIIH, a complex of ten proteins containing several structural domains with different activities. Christoph Meier (Oxford Protein Production Facility, UK) described some of the high-throughput techniques used in his laboratory, including limited proteolysis and protein surface engineering through chemical methods, such as the methylation of lysine residues. Darren Hart (EMBL, Grenoble, France) presented an automated high-technology method (ESPRIT, expression of soluble proteins by random incremental truncation), which allows one to test all the unidirectional truncations of a target protein (both N- and C-terminal) for soluble protein expression. Tobias Cornvik (Stockholm University, Sweden) described the CoFiblot method (colony filtration blot), which enables direct identification of soluble clones from large libraries of randomized constructs of the target gene and which was successfully applied to eukaryotic proteins where the N-terminal start point was randomized. Another experimental approach to improve the diffraction quality of the crystals was described by Zygmunt Derewenda (UVA School of Medicine, USA), who reported the results of a series of successful experiments of protein engineering on the surface residues involved in crystal packing interactions; this approach allows one to delineate an empirical and effective way to obtain better diffracting crystals.

4. Post-translational modifications

The biological information embedded in structural data cannot be extracted and properly evaluated if the possible post-translational modification of the proteins is disregarded. Moreover, it emerged during the discussions that post-translational modifications are often related to transitions between an ordered and a disordered conformational state of the protein and also to the expression system, which must account for them. Birgit and Frank Eisenhaber (Research Institute of Molecular Pathology – IMP, Vienna, Austria) reviewed the state of the art in post-translational modification predictions based on amino acid sequence, with particular emphasis on GPI lipid anchor sites, N-terminal N-myristoylation sites, farnesyl and geranylgeranyl anchor attachment, and the PTS1 peroxisomal signal. All these prediction methods are publicly available at http://annotator.org. Interestingly, the prediction rules appear to be independent of the phila. Another surprising, and experimentally verified, finding was the fact that some proteins carry sequence signals for post-translational modification or translocation that are silent in the normal biological context, but can become fully functional under specific conditions.

Animated discussions were focused especially on the interplay between computational and experimental approaches. In particular, several participants discussed the impact of post-translational modification on protein conformational disorder and on protein expression systems as well as the importance of some protein segments, which are known to be conformationally disordered though seem to be essential for protein expression and solubility. These are challenges for the bioinformatics community as well as for the experimentalists.

The success of the meeting suggests that it should be repeated in about two years to monitor the progress in the field and to improve further the interface between different scientific and technological approaches.

The financial support of the Bioinformatics Integration Network II (GAN-AU, Austria) is gratefully acknowledged. The workshop was the inaugural event organized by the newly appointed Training and Dissemination Center (Department of Biomolecular Structural Chemistry, University of Vienna, Austria) of the EU FP6 Integrated Project VIZIER (LSHG-CT-2004-511960).

Oliviero Carugo
Vienna University, Austria and University of Pavia, Italy

Kristina Djinovic Carugo
Vienna University, Austria

Alexander E. Gorbalenya
Leiden University Medical Center, The Netherlands

Paul Tucker
European Molecular Biology Laboratory, Hamburg, Germany