Large, dynamic, multi-protein complexes: a challenge for structural biology

2014 J. Phys.: Condens. Matter 26 463103

(http://iopscience.iop.org/0953-8984/26/46/463103)

J. Phys.: Condens. Matter **26** (2014) 463103 (11pp) doi:10.1088/0953-8984/26/46/463103

Topical Review

Large, dynamic, multi-protein complexes: a challenge for structural biology

Bartosz Różycki[1] **and Evzen Boura**[2]

[1] Institute of Physics, Polish Academy of Sciences, Al. Lotników 32/46, 02-668 Warsaw, Poland
[2] Institute of Organic Chemistry and Biochemistry AS CR, v.v.i., Flemingovo nam. 2., 166 10 Prague 6, Czech Republic

E-mail: rozycki@ifpan.edu.pl and boura@uochb.cas.cz

Received 8 August 2014, revised 8 September 2014
Accepted for publication 17 September 2014
Published 21 October 2014

Abstract

Structural biology elucidates atomic structures of macromolecules such as proteins, DNA, RNA, and their complexes to understand the basic mechanisms of their functions. Among proteins that pose the most difficult problems to current efforts are those which have several large domains connected by long, flexible polypeptide segments. Although abundant and critically important in biological cells, such proteins have proven intractable by conventional techniques. This gap has recently led to the advancement of hybrid methods that use state-of-the-art computational tools to combine complementary data from various high- and low-resolution experiments. In this review, we briefly discuss the individual experimental techniques to illustrate their strengths and limitations, and then focus on the use of hybrid methods in structural biology. We describe how representative structures of dynamic multi-protein complexes are obtained utilizing the EROS hybrid method that we have co-developed.

Keywords: protein structure, multi-protein complexes, hybrid methods of structural biology

(Some figures may appear in colour only in the online journal)

1. Introduction

Proteins are directly involved in practically all functions of biological cells, including transcription and translation of the genetic code, metabolism, signaling, transport and cell shaping. In fact, most events taking place in biological cells are either directly performed, regulated or catalyzed by proteins. This concerns the central dogma of molecular biology: DNA replication, DNA to RNA transcription, and translation of the genetic code from RNA into proteins. Proteins called transcription factors regulate DNA transcription. Other proteins such as protein kinases regulate transcription factors, and together with other signaling proteins comprise intricate self-regulating networks in living cells. Some proteins are metabolic enzymes that utilize various nutrients to produce energy, which is used to sustain vital processes in the cells. Other proteins mediate signal transduction, i.e., they control chains of biochemical events that allow the cell to respond to changes in its environment. A classic example is the production of second messengers (intracellular signaling molecules such as cyclic AMP) in response to extracellular stimuli (such as hormones or neurotransmitters). In addition, cells use proteins as construction material. The cytoskeleton—the cellular scaffolding present in every cell—consists entirely of proteins. Proteins also serve as pumps and channels in various cellular membranes. In these membranes, the weight ratio of lipids to proteins can even reach 1 : 1.

Proteins orchestrate the overwhelming majority of functions in any organism. With the pioneering work of John Kendrew and Max Perutz (Nobel Prize in Chemistry in 1962 for determining the first atomic structures of proteins using x-ray crystallography) it became evident that proteins are folded into appropriate native structures to perform their functions. Several decades later it appeared that this dogma is not true in general as some proteins are intrinsically disordered. Determination of the atomic structure of a macromolecule can often lead to the understanding of how the given molecule performs its biological functions. The canonical example is the structure of the DNA double helix, which explained the mechanism of DNA replication. The ever-increasing

0953-8984/14/463103+11$33.00

demand for bio-molecular structures led to the establishment of structural biology in 1950s and 1960s. Structural biology elucidates structures of bio-molecules such as proteins, nucleic acids, and their complexes to understand the mechanism of their function at atomic detail. Among the most striking of recent successes are the structures of G-protein coupled receptors [1, 2] that explain how signal transduction across the plasma membrane occurs, and how hormones and opiates act on their targets. Other prominent examples include the structures of entire viruses, such as the human pathogen Enterovirus 71, which help to explain the mechanism of infection and provide the basis for rational drug design [3]. A leading example of rational drug design is the work of Balbas and colleagues [4]. Here, molecular dynamics simulations provided the rationale for a focused chemical screen that identified compounds suppressing the growth of drug resistant prostate cancer cells. It is expected that rational drug design will become increasingly important in medicine [5].

Structural biology originated from x-ray crystallography. Today, a much broader spectrum of methods are exploited, as often problems cannot be solved by x-ray crystallography alone. Notable examples are intrinsically disordered proteins (IDPs) and multi-domain proteins in which individual domains are connected with long flexible linkers. In addition to x-ray crystallography, contemporary structural biology commonly uses such methods as nuclear magnetic resonance (NMR), cryo-electron microscopy (cryoEM), small angle x-ray scattering (SAXS), Förster resonance energy transfer (FRET), and electron paramagnetic resonance (EPR). These methods have similarly originated from condensed matter physics and have been gradually adapted to the needs of structural biology. In this review, we will discuss these individual methods only briefly and in the context of their strengths and weaknesses, as excellent textbooks and reviews are already available [6]. In particular, we will not discuss any of their physical principles and instead only describe their principal usage in contemporary structural biology. We will set up these techniques as a means to focus on hybrid methods that result from combining the plethora of available structural biology methods. Such hybrid methods are being rapidly developed [7–9] in an attempt to characterize large, dynamic protein complexes that have it all: many well-structured domains, long flexible linkers, and disordered segments that can adopt a unique structure only upon a biologically significant event such as complex formation. This molecular arrangement is often present in protein complexes that facilitate intracellular trafficking, such as the ESCRT or BLOC complexes [10, 11].

2. High-resolution methods: x-ray crystallography and NMR

X-ray crystallography and nuclear magnetic resonance (NMR) are considered the 'bread and butter' of structural biology as they provide elaborate atomic models of biomacromolecules. Recently, also cryo-electron microscopy (cryoEM) has been used to derive atomic models [12]. Here, however, the macromolecular systems of interest typically must have specific properties such as high symmetry to achieve atomic resolution, which makes the icosahedral viruses an ideal system for cryoEM studies [13]. Only very recently, due to technical advancements such as the introduction of direct detection device imaging cameras, have non-symmetric macromolecules been probed to atomic resolution in cryoEM experiments [14]. An excellent brief review of recent developments in cryoEM is provided in [15].

The experimental pipelines for x-ray crystallography and NMR are similar in the sense that pure and homogeneous protein samples are required, and that analysis of experimental data permits to solve protein structures with atomic resolution. Therefore, we will describe the crystallographic pipeline (see figure 1) and then clarify the differences in the protein NMR pipeline. First, the protein of interest has to be prepared in sufficient amount (usually a few milligrams suffice) and in high purity. Recombinant proteins are used almost exclusively. Genes are cloned into suitable plasmids and the proteins are usually obtained by over-expression in *E. coli*, yeast, insect or mammalian cells. Even though protein purification may be difficult and tedious in some cases, usually routine biochemical methods are sufficient. In the early days, the bottleneck in bio-molecular crystallography was data analysis and structure determination; today, the growth of crystals is the major obstacle in protein crystallography. A protein solution is mixed with a precipitant and the mixture drop is left to equilibrate—usually by vapor diffusion. With the introduction of protein crystallization robots, screening of thousands of crystal conditions has become commonly accessible. Once the crystals are obtained, diffraction data are collected either at a home source of x-rays or, more commonly, using a synchrotron source. The phase problem of molecular x-ray crystallography can be solved by incorporating a heavy element (crystal soaking by heavy elements, or the use of selenomethionine) or by applying molecular replacement, a technique that is been used with increased frequency. In the latter method, the initial phases are 'guessed', based on the atomic structure of a similar macromolecule. Analysis of the diffraction data usually requires just a few clicks in a crystallographic program package on a personal computer, with the exception of difficult cases: high non-crystallographic symmetry, poor diffraction profiles, radiation damage, high mosaicity, low resolution. For more details, excellent books can be consulted [16, 17].

Protein NMR, founded on totally different physical principles than x-ray crystallography, is also commonly used to solve atomic structures of proteins. The main difference is that x-ray crystallography requires protein crystals whereas protein NMR experiments are performed in solution but require labeled proteins. For NMR measurements the protein of interest must be labeled by stable isotopes (usually ^{13}C and ^{15}N), which can be challenging in the case of insect or mammalian expression system. The protein must be stable at the time of NMR measurement—which can last up two weeks—and must be relatively small (40–50 kDa usually), although gradual progress is being made toward measurements of ever larger proteins (reviewed in [18]). Here, the major obstacle is that the slow tumbling of large proteins in solution

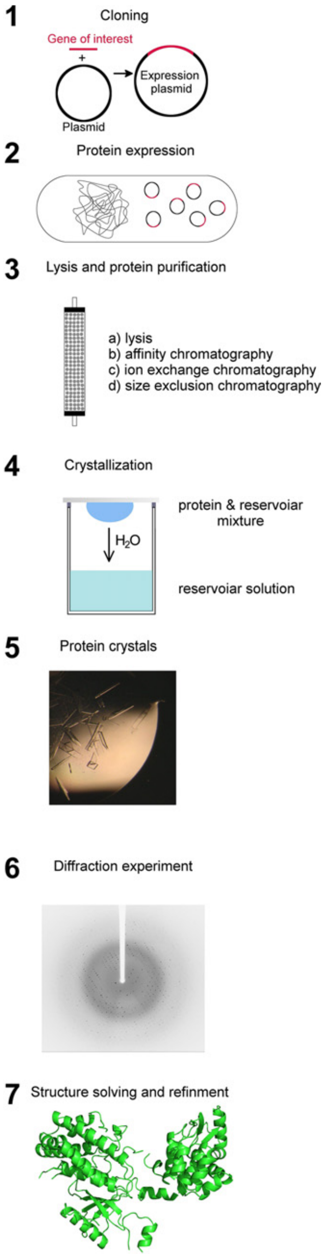
1
Cloning
Gene of interest
+
Plasmid
Expression plasmid
2
Protein expression
3
Lysis and protein purification
a) lysis
b) affinity chromatography
c) ion exchange chromatography
d) size exclusion chromatography
4
Crystallization
protein & reservoir mixture
H_2O
reservoiar solution
5
Protein crystals
6
Diffraction experiment
7
Structure solving and refinment

Figure 1. Protein crystallography pipeline. (1) Cloning—the gene coding the protein of interest is cloned in the expression plasmid for a selected organism. The bacteria *Escherichia coli* is used most often but for 'difficult targets' the use of insect or mammalian cells might be necessary. (2) Protein expression—the expression plasmid is transferred in the *E. coli* cells. The resulting genetically modified *E. coli* cells harbor (besides its own genomic DNA) the expression plasmids. (3) Lysis and protein purification—after the protein is expressed, the bacterial cells are lysed (broken down) and the protein is purified using several rounds of chromatographic techniques. (4) Crystallization trials—depicted is the most popular hanging drop method. In this setup the protein solution is mixed with the reservoir solution, and the drop equilibrates by diffusion against the reservoir solution. The concentration of the protein increases in time to the point where crystals might be obtained. (5) Protein crystals—usually several hundreds of different reservoir solutions are screened for the ability to induce crystallization. Shown is a 400 nl drop containing protein crystals. (6) Diffraction experiment—shown is a diffraction pattern of a protein crystal that was measured in a synchrotron facility. (7) Structure solving and refinement—structure of a lipid kinase solved by exactly this pipeline recently in our laboratory [84].

leads to a fast decay of the NMR signal. The recorded spectra are analyzed and the atomic model is built using specialized software, in a similar but usually more time-consuming way as in the case of protein crystallography. Details can be found in [19].

3. Limitations of x-ray crystallography and NMR

The necessary condition for the formation of diffracting crystals is self-organization of macromolecules into identical, repeating asymmetric units. As a consequence, practically all macromolecules in the crystal—or sets of macromolecules, if there is more than one macromolecule in the asymmetric unit—must adopt the same conformation (i.e. structural arrangement). This represents a problem because proteins must be able to change their conformations to fulfill their respective physiological functions. Usually, a particular protein conformation corresponds to a local free-energy minimum that is deep enough to form a stable conformation in the crystal. In more difficult cases, it may be necessary to stabilize a single conformation with stabilizing mutations, small molecule inhibitors, antibodies, or a combination of these. However, if the protein of interest is flexible—i.e. its free-energy landscape is rough, with many local minima separated by small barriers—it cannot be crystallized and other methods of structural biology must be used. For proteins with molecular weights below ∼100 kDa, the method of choice is solution NMR, which can yield structural ensembles and, thus, deal with molecular flexibility. For larger systems, however, hybrid methods must be used.

Not all proteins at physiological conditions are folded into functional three-dimensional native structures. Indeed, intrinsically disordered proteins (IDPs) resemble a flexible polymer chain and fold only under particular circumstances, for example, when they bind specifically to another biomolecule [20, 21]. Since IDPs are usually rather small, they are directly accessible to solution NMR measurements [22]. Another promising technique to explore their conformations is the use of advanced computational tools such as molecular dynamics simulations restrained by SAXS data [23, 24]. For more details, excellent reviews can be consulted [20, 21].

Many protein complexes that are currently in the focus of structural biology are neither completely folded into stable structures nor intrinsically disordered. For instance, a common architecture among protein complexes involved in intracellular trafficking is several well-folded domains connected with long flexible linkers that behave individually like IDPs, as in Endosomal Sorting Complexes Required for Transport (ESCRTs) [25, 26], Biogenesis of Lysosome-related Organelles Complexes (BLOCs) [11, 27] or Autophagy-related Genes (ATGs) [28]. It is usually extremely difficult to derive atomic models for such protein complexes: they are not directly accessible to x-ray crystallography due to the presence of highly disordered and dynamic segments (although their separate domains can be crystallized); they are also not accessible to solution NMR due to their large molecular weights, usually between 100 to 300 kDa—well beyond the capacity of contemporary NMR techniques; and their inherent flexibility and the lack of symmetries make them practically inaccessible to cryoEM, although some progress has been made recently [29]. Therefore, to elucidate representative conformations of such protein complexes, various low-resolution methods as well as combinations of low- and high-resolution methods must be employed [30–32]. In the sections below, we will discuss SAXS and selected spectroscopy methods that are based on site-directed labeling, and how they can be combined with high-resolution methods to derive atomic models.

4. SAXS

As discussed above, the high-resolution methods of structural biology have certain shortcomings: x-ray crystallography is optimally suited for well-folded proteins and tightly bound bio-molecular complexes whereas solution NMR is limited to moderate molecular sizes. SAXS offers a promising alternative for the structural characterization of proteins and bio-molecular complexes in solution. It is commonly regarded as a low-resolution method. In fact, the resolution of the SAXS method is inherently limited as an intricate three-dimensional molecular structure is reduced to a one-dimensional intensity profile (unlike x-ray crystallography data, the SAXS signal is spherically averaged). Despite the resulting loss in information, careful analysis and interpretation of the SAXS intensity profile can lead to deep insights into structure-function relationships [33, 34].

Sample preparation for SAXS experiments is practically identical as in the pipeline of bio-molecular crystallography up to the stage of crystal growth (i.e. pure and monodisperse sample), which makes the SAXS method particularly well suited to supplement crystallographic studies. In fact, unlike x-ray crystallography (which relies on a crystalline order to produce diffraction patterns), SAXS experiments always provide scattering data because macromolecules always scatter in solution. Furthermore, the SAXS signal does not broaden or attenuate when applied to dynamic or flexible systems. Unlike solution NMR techniques, SAXS is not limited by the molecular size. SAXS is therefore a robust technique that can be applied to a wide variety of solution conditions, molecular concentrations and temperatures. As such, SAXS is often used to characterize shapes and dimensions of proteins in solution [33, 34]. A standard approach is to use the scattering intensity profile to determine the pair−distance distribution function and the corresponding molecular envelope [35–37]. Such molecular envelopes provide informative visual interpretation of the observed SAXS data. However, when supplemented with some other structural information, SAXS data can be used much more efficiently. For example, SAXS can be used to determine structures of protein complexes if atomic structures of the constituent proteins are known. To achieve this goal with optimal accuracy, structural models of the protein complexes should be fitted directly to the experimental SAXS data; simply placing atomic models of the proteins into a molecular envelope does not fully use the structural information encoded in the scattering intensity profile.

The most widely distributed program for computing SAXS intensity, $I\,(q)$, from atomistic models is CRYSOL [38]. Its direct application is to verify whether a particular crystal structure can faithfully represent the protein thermodynamic state in solution. CRYSOL-related software such as SASREF and CORAL [37, 39] are routinely used to determine the correct oligomeric state of protein complexes. This goal is typically achieved by selecting an optimal model out of a pool of models generated by rigid-body docking. Structure refinement based solely on SAXS data is rare and challenging. However, reconciling the solution and crystal states can be achieved with the help of normal mode analysis [40]. Also here, SAXS data are often used as a filter to identify the correct atomistic model. An alternative approach is to integrate the model−data discrepancy as a pseudo-potential function into the model refinement procedure [41, 42].

A clear advantage to using SAXS data in molecular modeling is that SAXS experiments are performed in aqueous environments and, thus, provide information about the thermodynamic state of molecules in solution. On the other hand, the SAXS intensity profile must be taken as a difference in signals between the sample and the corresponding buffer, which may lead to significant systematic errors if the signal subtraction is inadequate. Also, the hydration shell on the protein surface must be considered in SAXS modeling. We note that different programs for computing SAXS intensity profiles from atomic models (such as CRYSOL [38], FoXS [43], AXES [44], AquaSAXS [45] and SASTBX [46]) differ in how they treat the hydration shell, which puts additional uncertainty on SAXS-derived models.

Many proteins that are currently under the focus in molecular biology are flexible to some degree. Aforementioned examples are IDPs and multi-domain proteins with flexible segments. Identifying flexibility from SAXS data is often deduced from the Kratky plot (i.e. the plot of $q^2 I\,(q)$ as a function of the momentum transfer q, see the left-hand panels of figure 2). Convergence of the Kratky plot at high q suggests compaction, whereas a hyperbolic shape suggests flexibility (compare the upper and lower panels on the left-hand side of figure 2). The hyperbolic feature is a trademark of random coils and IDPs. However, the Kratky interpretation may be difficult to assess if the SAXS data are noisy or truncated. Recently, Rabmo and colleagues introduced the use of the Porod−Debye law (analysis of $q^4 I\,(q)$ versus q^4 at intermediate q-values, see the right-hand panels of figure 2) as a more robust approach to distinguish between rigid and flexible systems [47]. In addition, molecular flexibility can be presumed if SAXS data cannot be explained with a single model, suggesting that an ensemble of models may be required to account for the experimental data [25, 48]. We note that in addition to the analysis of SAXS data, it is critical to evaluate molecular flexibility using biochemical or biophysical methods such as limited proteolysis or hydrogen/deuterium exchange.

Flexible protein systems require ensemble-modeling strategies that attempt to use multiple structural models to fit experimental data. There might also be situations in which the thermodynamic state of macromolecules in solution consists of several distinct, compact conformations comprising

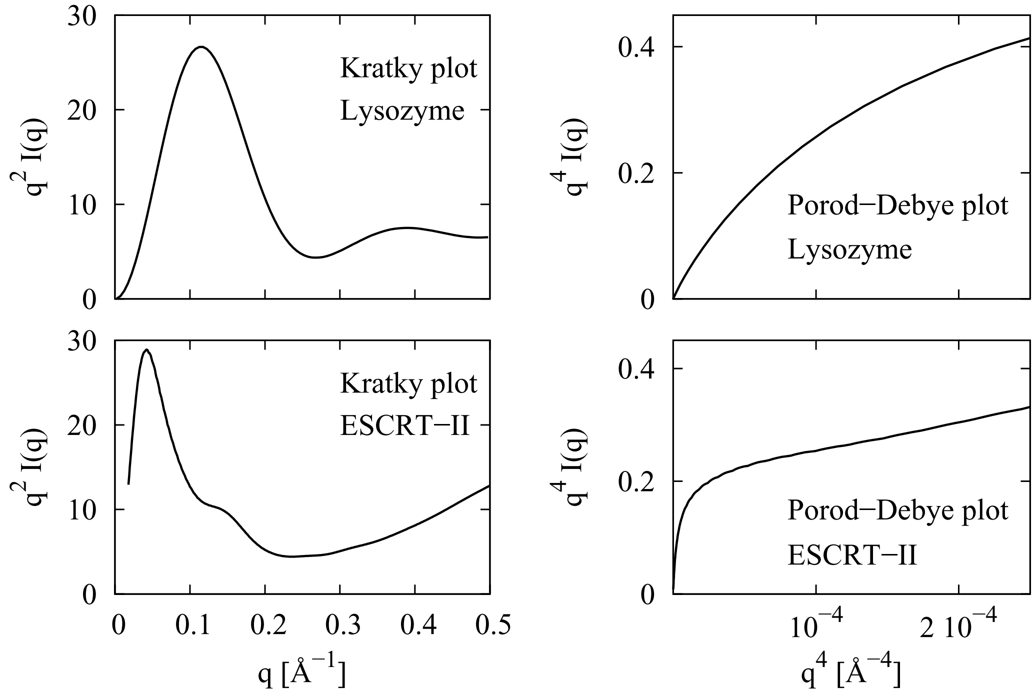

Figure 2. SAXS data of lysozyme (upper panels) and ESCRT-II complex (lower panels) in the Kratky (left-hand panels) and Porod–Debye (right-hand panels) representations. Lysozyme is a compact protein with a unique native structure whereas ESCRT-II is a multi-domain protein complex containing a number of disordered segments and flexible loops. The flexibility in the ESCRT-II proteins can be deduced both from the Kratky plot at $q > 0.3$ Å^{-1} and from the Porod–Debye plot at $q \approx 0.1$ Å^{-1}, see main text. The lysozyme SAXS data are provided as an example in the CRYSOL package [38]. The ESCRT-II SAXS data are taken from our recent work [52].

a 'nonflexible ensemble'. Examples are transient complexes that have been observed in weakly interacting proteins [49]. Several SAXS-based approaches for modeling structural ensembles have been developed, including the ensemble optimization method (EOM) [30], the SAXS module in the integrative modeling platform [50], the minimal ensemble search [51], and the ensemble refinement of the SAXS (EROS) method [8]. All these approaches require a large pool of protein conformations as an input. The conformation pool can be created based on steric exclusion [30], high-temperature molecular dynamic simulations [51], statistical potentials for protein binding [8, 50], or topology-based Go-type models [31]. After the pool has been generated, heuristic algorithms are usually employed to determine which combination of conformations best fits the SAXS data. In contrast, the EROS method uses a different strategy in which the pool of simulation structures is only gently reweighed to improve the agreement with the SAXS data. To refine the simulation ensemble in a controlled way, and to prevent data over-fitting, the maximum-entropy method is used. The EROS method has been further developed to combine SAXS with spectroscopy experiments based on site-directed labeling [25, 52].

5. Spectroscopy methods based on site-directed labeling

Contemporary structural biology uses two distinct spectroscopy methods that rely on site-directed labeling: Förster resonance energy transfer (FRET), and electron paramagnetic resonance (EPR). The physical principles of these techniques are different, and yet their usage in structural biology is similar. Both FRET and EPR spectra can provide useful information on a variety of properties, including rotational and vibrational degrees of freedom, micro-environment of the labels, their dynamics and many others [53–55]. However, their primary application in structural biology exploits their common feature—that both methods can be used to obtain information about inter-label distances. In one case, distances between molecular electrical dipoles, and in the other, between unpaired electrons. A set of such distances provides structural information. In fact, if the set of distances is large enough, e.g. 40 distances for a small protein domain, one can even solve the structure of the domain, not unlike in protein NMR. Indeed, there are recent cases where protein structures have been solved using only FRET or EPR [56–59]. But much more often these techniques are used to monitor protein folding [60, 61], protein conformational changes [62], formation of bio-molecular complexes [63, 64], or to help elucidate the representative conformations of flexible protein assemblies [25, 52, 65]; more details can be found in recent excellent reviews [66, 67].

However, before using these spectroscopy methods, the macromolecules of interest have to be labeled at specific sites to permit structural interpretation of the spectroscopic data. The reactive amino groups present in lysine, arginine, asparagine or glutamine residues could be used for the purpose of molecular labeling, however, these four amino acids are too abundant even in small proteins. Instead, labels are attached to cysteines, which are the only amino acids possessing the reactive thiol group. In the case of proteins that have multiple surface-exposed cysteine residues, the cysteines are replaced ('mutated

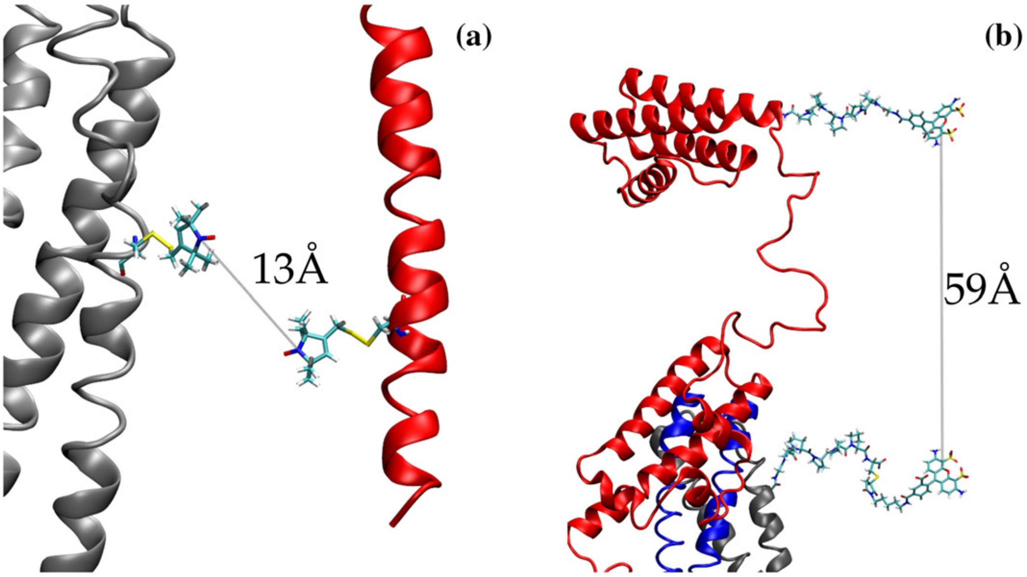

Figure 3. Site-directed labeling of proteins. (*a*) Spin labels (MTSL) are attached to two separate domains of a protein. Here, the spin labels are shown in the stick representation. The protein domains are shown in gray and red. DEER can be used to determine the distribution of distances between the active sites in the spin labels. The MTSL labels are attached to surface-exposed cysteine residues via disulfide bonds. (*b*) Fluorescence labels (Alexa448 and Alexa594) are attached to different domains of a protein complex. The fluorescence labels are shown in the stick representation. Note that the linkers of the fluorescence labels are much longer than the MTSL linkers shown in (*a*). The protein complex is shown in the cartoon representation, each protein in a different color. FRET can be used to determine distances between the active sites in the fluorescence labels. Thiol click chemistry is used to attach the fluorescence labels to cysteine residues on the surface of the protein complex.

out') by chemically similar non-reactive amino acid residues like serine or alanine. Such mutations are implemented using standard methods of modern molecular biology: the DNA encoding the protein of interest is changed and subsequently the modified protein is expressed. Unique pairs of cysteine residues can be placed in almost any desired place on the protein surface (e.g. close to a binding site), and thiol click chemistry or disulfide bond formation might be used to attach fluorescent or spin labels to them (see figure 3). For the EPR method, the commercially available small paramagnetic MTSL label (see figure 3(*a*)) is almost exclusively used and attached to cysteine residues via disulfide bonds. In the case of many commercially available fluorescent labels (see figure 3(*b*)), click chemistry based on reaction between a maleimide group with the cysteine thiol is usually used [68]. The advantage is that the resulting bond is stable in reducing environments unlike disulfide bonds. On the other hand, the fluorescence label linker is much longer than the MTSL linker (compare the structures of the fluorescence and spin labels shown in figure 3), which usually does not pose a problem as FRET is used to measure larger distances, practically up to two Förster radii, which for some dye pairs might be up to 200 Å.

An inherent difficulty in FRET experiments is that two different dyes (donor and acceptor) must be used. To achieve a good signal-to-noise ratio, often a stochastic labeling approach is sufficient: The bio-molecule of interest is labeled by a small amount of donor and a large amount of acceptor, which, due to the binomial distribution, leads to many acceptor–acceptor labeled bio-molecules (that are 'invisible' in FRET measurements), some amount of the desired donor–acceptor labeled bio-molecules and only a small number of donor–donor labeled bio-molecules. The signal arising from donor–donor labeled bio-molecules may be more or less subtracted.

In general, there are two categories of FRET experiments. Bulk FRET provides information only about the average inter-label distance but can be performed in relatively high concentrations, which is important for experiments on transient protein complexes with micromolar dissociation constants. Single molecule FRET (smFRET) also provides information about variations in inter-label distances but requires that only one molecule is measured at a time. This requirement can be achieved by using very low concentrations of labeled molecules and minimizing the volume in which measurements are performed (confocal microscopes). Alternatively, methods to isolate single molecules to a microscope field of view are utilized. For example, the tethering of DNA to a coverslip has been utilized to specifically observe the dynamics and structural basis of protein−DNA interactions by microscopy [69]. The outcome of a smFRET experiment is a histogram depicting molecular states or conformations. The simplest analysis is by Gaussian fitting, but more reliable methods have been developed, including direct comparison with molecular dynamics simulations [70–72].

A pulsed EPR experiment that can provide information about distances between spin labels is the double electron-electron resonance (DEER) experiment. Depending on the experimental setup and protein concentration (generally, the higher concentrations the better), the method is sensitive to up to 50 Å but with some modification it can be used to measure distances even up to 80 or 90 Å [73]. DEER and FRET data are similar in that the inter-label distance distributions can be deduced from both. One inherent difficulty

Table 1. Summary of the structural biology methods discussed in this review article.

Method	Requirements and Limitations	Advantages
X-ray crystallography	– x-ray source – well-ordered proteins that form crystals – flexible proteins cannot be resolved – 'phase problem'	– accurate atomic structures – versatile method; applicable to systems ranging from small domains to huge protein assemblies – easy model building with user-friendly open source software – commonly accessible synchrotron facilities
Protein NMR	– expensive NMR spectrometers – ^{13}C and ^{15}N labeled proteins –limited to small and medium size proteins – resonance assignment and model building is often not straightforward	– provides high-resolution models of proteins in solution – no need for protein crystals – applicable to 'flexible macromolecules' such as IDPs – can provide information about protein dynamics
cryoEM	– very expensive electron microscopes – limited to large proteins and protein assemblies – state-of-the-art algorithms for image analysis are computationally very demanding	– suitable for analysis of large protein assemblies that are difficult for other methods – very small amount of sample needed – no need for protein crystals – exploits internal symmetries of protein systems such as virus capsids
SAXS	– x-ray source – spherically averaged scattering signal implies low-resolution – cannot be used to derive atomic structures	– simple sample preparation (unlabeled proteins in solution) – high-throughput screening potential – fast data analysis – beautifully complements x-ray, NMR and other methods – many synchrotron facilities have dedicated bio-SAXS beamlines
FRET and DEER	– proteins labeled with fluorescent or paramagnetic probes – usually mutants with cysteine residues at desired places must be prepared – difficult to use to derive atomic structures	– direct observation of flexibility and/or of conformational changes – no limitations on protein size – applicable to membrane proteins – single-molecule FRET experiments require extremely small amounts of protein sample

in the DEER method, similarly as in SAXS, is that incorrect background subtraction may lead to wrong interpretation of the data. Another issue arises when deriving inter-label distance distribution from the dipolar evolution function. A Gaussian distance distribution is often assumed but a much more reliable approach is modeling based on MTSL conformation library [74].

Both FRET and EPR have certain limitations and disadvantages but nevertheless these are powerful techniques in modern structural biology. When combined with SAXS and/or high-resolution methods, into hybrid methods, FRET and EPR become critically important tools.

6. Hybrid methods

All methods of structural biology have limited scope and applicability, as summarized in table 1. In difficult cases, none of these methods can provide the desired bio-molecule structures. However, if two or more of them are combined, they can often supplement each other to produce accurate models of bio-molecules and their assemblies. For example, cryoEM-derived maps have been used in protein crystallography as search models for molecular replacement to obtain atomic models [75, 76]. As already discussed in section 4, protein crystallography is increasingly complemented by SAXS to determine structures of multi-domain proteins and protein complexes [7, 28, 77]. Hybrid methods combining NMR with SAXS can be used to refine high-resolution structures of proteins in aqueous environments [42]. The latter method is particularly powerful because NMR and SAXS provide complementary information: NMR data impose constrains on distances (NOE) or dihedral angles (J-coupling) while SAXS data provide information on the overall size and shape of the macromolecule. Solution NMR and SAXS can also be combined to determine conformations of dynamic protein complexes with flexible linkers [78]. In the latter approach, chemical shift perturbations provide information about the binding modes, and SAXS helps to position the constituent proteins relative to each other.

Particularly powerful hybrid methods in determining conformations of large, dynamic, multi-protein complexes are those which combine x-ray crystallography or NMR (which provide high resolution structures of individual domains) with SAXS (which gives information on the global size and shape of the molecular assembly) with DEER or FRET (which impose local restrains on distances between selected sites). In fact, with the help of methods like EROS, data from x-ray crystallography, NMR, SAXS, DEER and FRET experiments can be combined and used together to obtain detailed representations of the structures and motions in systems ranging from the ESCRT membrane-protein trafficking system [25, 52] to protein kinases in dynamic complexes with phosphatases [48, 78].

We now illustrate how hybrid structures are obtained in the EROS method that we co-developed (see figure 4). The EROS method proceeds in two steps: first, a coarse-grained model for protein binding is used to simulate the protein system under study and, in this way, an initial ensemble of protein

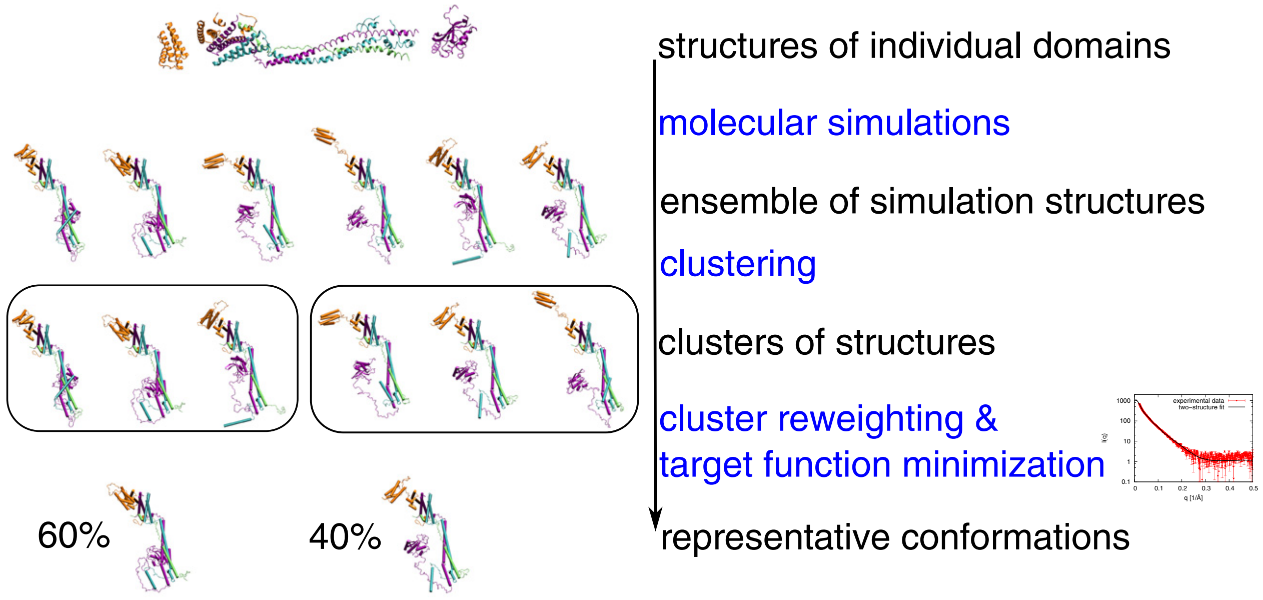

Figure 4. Ensemble refinement procedures. Structures of individual domains of a multi-protein complex are used as input for molecular simulations. The resulting conformations of the protein complex are clustered according to their structure similarity. The clusters are next reweighted to minimize a target function. The resulting ensemble of conformations fits experimental data (SAXS, DEER and/or FRET). Representative conformations of the protein complex can be inspected visually for possible interpretation.

configurations is generated. Second, the simulation ensemble is refined to improve agreement with experimental data. Within the framework of our coarse-grained model, folded protein domains are represented as rigid bodies (based on their atomic structures) whereas disordered loops and flexible linkers connecting the domains are represented as chains of amino acid beads with appropriate stretching, bending, and torsion-angle potentials [79]. The interactions between the domains are described at the residue level with statistical amino-acid-dependent potentials and Debye–Hückel-type electrostatics. By changing the Debye length, which controls the range of the effective electrostatic interactions in the simulations, different buffer ionic strengths can be captured. The transferable energy function used in the EROS simulations has been shown to correctly predict structures and binding affinities of protein complexes [79]. This result indicates that our simulation model properly samples the relevant conformational space of protein complexes.

The protein configurations generated in the simulations are next segregated into clusters based on their structure similarity, and statistical weights are assigned to the clusters. Before any refinement, these statistical weights are simply proportional to cluster populations. The computed, ensemble-averaged quantities such as SAXS intensity profiles, EPR dipolar evolution functions and FRET efficiency histograms, which should be compared directly to experimental data, are functions of the cluster weights. In the course of the ensemble refinement, the relative weights of the clusters are varied to improve agreement with experimental data. To prevent data over-fitting, a minimum entropy method is used. This method is based on minimization of a pseudo-potential function that consists of a model−data discrepancy function and a cross-entropy term that quantifies how different the refined ensemble is from the original simulation ensemble. Minimizing this pseudo-potential function leads usually to gentle reweighting of the clusters only, which reflects our confidence in both experiments and simulations.

Another way of refining the simulation ensemble is the minimum ensemble method that selects the smallest possible set of structure clusters that accounts for experimental data. The advantage of this method is that it usually produces only a small set of representative structures that can be easily inspected visually. However, by discarding a significant portion of the simulation ensemble, the minimum ensemble method does not fully exploit the predictive power of molecular simulations.

In principle, the structural ensemble can be fitted either to raw experimental data or to commensurate quantities such as SAXS-derived pair−distance distribution function or DEER-derived inter-label distance distribution. However, to avoid introducing any regularization-dependent artifacts into the ensemble refinement, the simulation structures are fitted directly to experimental data in the EROS method. We also note that a sensible practice is to cross-validate the structural ensemble with independent datasets excluded from refinement.

All methods of structure determination require molecular modeling. Determination of protein structures from NMR or crystallographic data certainly requires all-atom modeling. Coarse-grained approaches are sufficient to structurally interpret SAXS data because of a relatively low resolution of the SAXS method (usually far below the size of single amino acids). For the spectroscopy methods based on site-directed labeling of proteins, atom-level modeling of the labels is crucial to correctly interpret the experimental spectra. Since the EROS method has been developed to study large multi-domain protein complexes that can undergo vast conformational fluctuations, it uses efficient coarse-grained simulations to sample physical configurations of the proteins.

However, all-atom structures of the spin and fluorescence labels are used in the ensemble refinement procedure to compare the simulation output to DEER and FRET data, respectively.

7. Summary and outlook

Among the proteins that are most difficult to characterize structurally are those which have several large, well-folded domains connected with long flexible linkers. Although abundant and critically important in cell physiology, such proteins have been intractable by regular methods of structural biology. In fact, their large size coupled with the flexibility of the linkers mean there is currently no single technique that can provide information on the overall structure. However, several distinct techniques—protein crystallography, NMR, SAXS, EPR, FRET—combined together with the help of advanced computational tools can resolve the dominant conformations of the large, flexible protein assemblies.

Among the first multi-domain, flexible protein complexes characterized by the hybrid methods were the ESCRT complexes. We expect that analogous hybrid methods will be applied to such flexible protein systems as BLOC and ATG complexes in the near future. Application of hybrid methods in structural biology should be preceded by the appropriate development of suitable open-source software with a user-friendly interface. In fact, advanced software for structure determination and refinement is indispensible in the fields of protein crystallography [80], NMR [81], SAXS [37] and—perhaps to less extent—DEER [74] and FRET [59]. The development of user-friendly computer programs for ensemble refinement with multiple datasets would be a considerable service to the structural biology community.

We expect that a pivotal direction in structural biology will be in the development of new methods to study conformations of membrane-bound and transmembrane proteins. Significant results were obtained using smFRET in combination with molecular dynamics simulations [82]. However, small angle neutron scattering (SANS) might become as a powerful tool to study membrane proteins as SAXS is for soluble proteins. The inherent advantage of SANS is that under appropriate conditions lipids scatter neutrons differently than proteins do [83]. This feature may permit comparison of solution and membrane-bound states of membrane binding proteins, or direct observation of different conformations of transmembrane proteins (e.g. a transmembrane receptor with and without its ligand).

Acknowledgments

The research of BR is supported by the Polish National Science Center Grant No. 2012/05/B/NZ1/00631. The research in the laboratory of EB is supported by MarieCurie FP7-PEOPLE-2012-CIG project number 333916, by Project InterBioMed LO1302 from Ministry of Education of the Czech Republic, and by the Academy of Sciences Czech Republic (RVO: 61388963). We thank Grzegorz Nawrocki, Mateusz Chwastyk, Michael Downey and David Murray for critical reading of this review.

References

[1] Kruse A C *et al* 2013 Activation and allosteric modulation of a muscarinic acetylcholine receptor *Nature* **504** 101–6

[2] White J F *et al* 2012 Structure of the agonist-bound neurotensin receptor *Nature* **490** 508–13

[3] Plevka P *et al* 2012 Crystal structure of human enterovirus 71 *Science* **336** 1274

[4] Balbas M D *et al* 2013 Overcoming mutation-based resistance to antiandrogens with rational drug design *eLife* **2** e00499

[5] Winter A *et al* 2012 Biophysical and computational fragment-based approaches to targeting protein–protein interactions: applications in structure-guided drug discovery *Q. Rev. Biophys.* **45** 383–426

[6] Liljas A 2009 *Textbook of Structural Biology* (Hackensack, NJ: World Scientific) p 572

[7] Rambo R P and Tainer J A 2013 Super-resolution in solution x-ray scattering and its applications to structural systems biology *Annu. Rev. Biophys.* **42** 415–41

[8] Rozycki B, Kim Y C and Hummer G 2011 SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions *Structure* **19** 109–16

[9] Yang S and Roux B 2011 EROS: Better than SAXS! *Structure* **19** 3–4

[10] Rozycki B *et al* 2012 Membrane-elasticity model of Coatless vesicle budding induced by ESCRT complexes *PLoS Comput. Biol.* **8** e1002736

[11] Lee H H *et al* 2012 Assembly and architecture of biogenesis of lysosome-related organelles complex-1 (BLOC-1) *J. Biol. Chem.* **287** 5882–90

[12] Bai X C *et al* 2013 Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles *eLife* **2** e00461

[13] Grigorieff N and Harrison S C 2011 Near-atomic resolution reconstructions of icosahedral viruses from electron cryo-microscopy *Curr. Opin. Struct. Biol.* **21** 265–73

[14] Lu P *et al* 2014 Three-dimensional structure of human gamma-secretase *Nature* **512** 166–70

[15] Smith M T and Rubinstein J L 2014 Structural biology. Beyond blob-ology *Science* **345** 617–9

[16] Rhodes G 2006 *Crystallography Made Crystal Clear: a Guide for users of Macromolecular Models* (*Complementary Science Series*) 3rd edn (Amsterdam/Boston, MA: Elsevier/Academic) p 306

[17] Woolfson M M1997 *An Introduction to X-ray Crystallography* 2nd edn (Cambridge: Cambridge University Press) p 402

[18] Frueh D P *et al* 2013 NMR methods for structural studies of large monomeric and multimeric proteins *Curr. Opin. Struct. Biol.* **23** 734–9

[19] Keeler J 2010 *Understanding NMR Spectroscopy* 2nd edn (Chichester: Wiley) p 511

[20] Fisher C K and Stultz C M 2011 Constructing ensembles for intrinsically disordered proteins *Curr. Opin. Struct. Biol.* **21** 426–31

[21] Rezaei-Ghaleh N, Blackledge M and Zweckstetter M 2012 Intrinsically disordered proteins: from sequence and conformational properties toward drug discovery *ChemBioChem* **13** 930–50

[22] Jensen M R, Ruigrok R W H and Blackledge M 2013 Describing intrinsically disordered proteins at atomic resolution by NMR *Curr. Opin. Struct. Biol.* **23** 426–35

[23] Bernado P and Svergun D I 2012 Analysis of intrinsically disordered proteins by small-angle x-ray scattering *Methods Mol. Biol.* **896** 107–22

[24] Daughdrill G W *et al* 2012 Understanding the structural ensembles of a highly extended disordered protein *Mol. Biosyst.* **8** 308–19

[25] Boura E *et al* 2011 Solution structure of the ESCRT-I complex by small-angle x-ray scattering, EPR, and FRET spectroscopy *Proc. Natl Acad. Sci. USA* **108** 9437–42

[26] Boura E *et al* 2012 Endosomal sorting complex required for transport (ESCRT) complexes induce phase-separated microdomains in supported lipid bilayers *J. Biol. Chem.* **287** 28144–51

[27] Kloer D P *et al* 2010 Assembly of the biogenesis of lysosome-related organelles complex-3 (BLOC-3) and its interaction with Rab9 *J. Biol. Chem.* **285** 7794–804

[28] Ragusa M J, Stanley R E and Hurley J H 2012 Architecture of the Atg17 complex as a scaffold for autophagosome biogenesis *Cell* **151** 1501–12

[29] Cossio P and Hummer G 2013 Bayesian analysis of individual electron microscopy images: towards structures of dynamic and heterogeneous biomolecular assemblies *J. Struct. Biol.* **184** 427–37

[30] Bernado P *et al* 2007 Structural characterization of flexible proteins using small-angle x-ray scattering *J. Am. Chem. Soc.* **129** 5656–64

[31] Yang S *et al* 2010 Multidomain assembled states of Hck tyrosine kinase in solution *Proc. Natl Acad. Sci. USA* **107** 15757–62

[32] Evrard G *et al* 2011 *D*ADIMODO: a program for refining the structure of multidomain proteins and complexes against small-angle scattering data and NMR-derived restraints *J. Appl. Crystallogr.* **44** 1264–71

[33] Graewert M A and Svergun D I 2013 Impact and progress in small and wide angle x-ray scattering (SAXS and WAXS) *Curr. Opin. Struct. Biol.* **23** 748–54

[34] Blanchet C E and Svergun D I 2013 Small-angle x-ray scattering on biological macromolecules and nanocomposites in solution *Annu. Rev. Phys. Chem.* **64** 37–54

[35] Franke D and Svergun D I 2009 DAMMIF, a program for rapid *ab initio* shape determination in small-angle scattering *J. Appl. Crystallogr.* **42** 342–6

[36] Svergun D I, Petoukhov M V and Koch M H J 2001 Determination of domain structure of proteins from x-ray solution scattering *Biophys. J.* **80** 2946–53

[37] Petoukhov M V *et al* 2012 New developments in the ATSAS program package for small-angle scattering data analysis *J. Appl. Crystallogr.* **45** 342–50

[38] Svergun D, Barberato C and Koch M H J 1995 CRYSOL—a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates *J. Appl. Crystallogr.* **28** 768–73

[39] Petoukhov M V and Svergun D I 2005 Global rigid body modeling of macromolecular complexes against small-angle scattering data *Biophys. J.* **89** 1237–50

[40] Stoddard C D *et al* 2010 Free state conformational sampling of the SAM-I riboswitch aptamer domain *Structure* **18** 787–97

[41] Zheng W J and Tekpinar M 2011 Accurate flexible fitting of high-resolution protein structures to small-angle x-ray scattering data using a coarse-grained model with implicit hydration shell *Biophys. J.* **101** 2981–91

[42] Grishaev A *et al* 2008 Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints *J. Biomol. NMR* **40** 95–106

[43] Schneidman-Duhovny D, Hammel M and Sali A 2010 FoXS: a web server for rapid computation and fitting of SAXS profiles *Nucl. Acids Res.* **38** W540–4

[44] Grishaev A *et al* 2010 Improved fitting of solution x-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling *J. Am. Chem. Soc.* **132** 15484–6

[45] Poitevin F *et al* 2011 AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformaly hydrated atomic models *Nucl. Acids Res.* **39** W184–9

[46] Liu H G, Hexemer A and Zwart P H 2012 The small angle scattering ToolBox (SASTBX): an open-source software for biomolecular small-angle scattering *J. Appl. Crystallogr.* **45** 587–93

[47] Rambo R P and Tainer J A 2011 Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod–Debye law *Biopolymers* **95** 559–71

[48] Francis D M *et al* 2011 Resting and active states of the ERK2:HePTP complex *J. Am. Chem. Soc.* **133** 17138–41

[49] Kim Y C *et al* 2008 Replica exchange simulations of transient encounter complexes in protein–protein association *Proc. Natl Acad. Sci. USA* **105** 12855–60

[50] Alber F *et al* 2008 Integrating diverse data for structure determination of macromolecular assemblies *Annu. Rev. Biochem.* **77** 443–77

[51] Pelikan M, Hura G L and Hammel M 2009 Structure and flexibility within proteins as identified through small angle x-ray scattering *Gen. Physiol. Biophys.* **28** 174–89

[52] Boura E *et al* 2012 Solution structure of the ESCRT-I and -II supercomplex: implications for membrane budding and scission *Structure* **20** 874–86

[53] Nesmelov Y E and Thomas D D 2010 Protein structural dynamics revealed by site-directed spin labeling and multifrequency EPR *Biophys. Rev.* **2** 91–9

[54] Dosremedios C G and Moens P D 1995 Fluorescence resonance energy-transfer spectroscopy is a reliable ruler for measuring structural-changes in proteins—dispelling the problem of the unknown orientation factor *J. Struct. Biol.* **115** 175–85

[55] Taraska J W 2012 Mapping membrane protein structure with fluorescence *Curr. Opin. Struct. Biol.* **22** 507–13

[56] Brunger A T *et al* 2011 Three-dimensional molecular modeling with single molecule FRET *J. Struct. Biol.* **173** 497–505

[57] Jao C C *et al* 2008 Structure of membrane-bound alpha-synuclein from site-directed spin labeling and computational refinement *Proc. Natl Acad. Sci. USA* **105** 19666–71

[58] Islam S M *et al* 2013 Structural refinement from restrained-ensemble simulations based on EPR/DEER data: application to T4 lysozyme *J. Phys. Chem.* B **117** 4740–54

[59] Kalinin S *et al* 2012 A toolkit and benchmark study for FRET-restrained high-precision structural modeling *Natural Methods* **9** 1218–25

[60] Chung H S *et al* 2012 Single-molecule fluorescence experiments determine protein folding transition path times *Science* **335** 981–4

[61] Chung H S and Eaton W A 2013 Single-molecule fluorescence probes dynamics of barrier crossing *Nature* **502** 685–8

[62] Rezabkova L *et al* 2010 14-3-3 protein interacts with and affects the structure of RGS domain of regulator of G protein signaling 3 (RGS3) *J. Struct. Biol.* **170** 451–61

[63] Lam A J *et al* 2012 Improving FRET dynamic range with bright green and red fluorescent proteins *Natural Methods* **9** 1005–12

[64] Boura E *et al* 2007 Both the N-terminal loop and wing W2 of the forkhead domain of transcription factor Foxo4 are important for DNA binding *J. Biol. Chem.* **282** 8265–75

[65] Caron N S *et al* 2013 Polyglutamine domain flexibility mediates the proximity between flanking sequences in huntingtin *Proc. Natl Acad. Sci. USA* **110** 14610–5

[66] Klare J P 2013 Site-directed spin labeling EPR spectroscopy in protein research *Biol. Chem.* **394** 1281–300

[67] Kalinin S *et al* 2010 Detection of structural dynamics by FRET: a photon distribution and fluorescence lifetime analysis of systems with multiple states *J. Phys. Chem.* B **114** 7983–95

[68] Ha T and Tinnefeld P 2012 Photophysics of fluorescent probes for single-molecule biophysics and super-resolution imaging *Annu. Rev. Phys. Chem.* **63** 595–617

[69] Kim H *et al* 2014 Protein-guided RNA dynamics during early ribosome assembly *Nature* **506** 334–8

[70] McKinney S A, Joo C and Ha T 2006 Analysis of single-molecule FRET trajectories using hidden Markov modeling *Biophys. J.* **91** 1941–51

[71] Best R B *et al* 2007 Effect of flexibility and cis residues in single-molecule FRET studies of polyproline *Proc. Natl Acad. Sci. USA* **104** 18964–9

[72] Merchant K A *et al* 2007 Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations *Proc. Natl Acad. Sci. USA* **104** 1528–33

[73] Jeschke G *et al* 2004 Sensitivity enhancement in pulse EPR distance measurements *J. Magn. Reson.* **169** 1–12

[74] Polyhach Y, Bordignon E and Jeschke G 2011 Rotamer libraries of spin labelled cysteines for protein studies *Phys. Chem. Chem. Phys.* **13** 2356–66

[75] Nemecek D, Plevka P and Boura E 2013 Using cryoEM reconstruction and phase extension to determine crystal structure of bacteriophage varphi6 major capsid protein *Protein J.* **32** 635–40

[76] Nemecek D *et al* 2013 Subunit folds and maturation pathway of a dsRNA virus capsid *Structure* **21** 1374–83

[77] Hammel M *et al* 2005 Structural basis of cellulosome efficiency explored by small angle x-ray scattering *J. Biol. Chem.* **280** 38562–8

[78] Francis D M *et al* 2011 Structural basis of p38alpha regulation by hematopoietic tyrosine phosphatase *Nature Chem. Biol.* **7** 916–24

[79] Kim Y C and Hummer G 2008 Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding *J. Mol. Biol.* **375** 1416–33

[80] Minor W *et al* 2006 HKL-3000: the integration of data reduction and structure solution–from diffraction images to an initial model in minutes *Acta Crystallogr.* D **62** 859–66

[81] Schwieters C D, Kuszewski J J and Clore G M 2006 Using Xplor-NIH for NMR molecular structure determination *Prog. Nucl. Magn. Reson. Spectrosc.* **48** 47–62

[82] Wang Y *et al* 2014 Single molecule FRET reveals pore size and opening mechanism of a mechano-sensitive ion channel *eLife* **3** e01834

[83] Clifton L A, Neylon C and Lakey J H 2013 Examining protein–lipid complexes using neutron scattering *Methods Mol. Biol.* **974** 119–50

[84] Baumlova A *et al* 2014 The crystal structure of the phosphatidylinositol 4-kinase II-alpha *EMBO Reports* **15** 1085–92