Exploring the spatial and temporal organization of a cell’s proteome

Martin Becka, Maya Toppb, Zachary Frazierc, Hariateo Tjongo, Min Xuc, Shiuhua Zhangc, Frank Alberc,⁎

a European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany
b Molecular Biology, Crystallography, Department of Biological Sciences, Birkbeck College, University of London, London, UK
c Program in Molecular and Computational Biology, University of Southern California, 1050 Childs Way, RRI 413E, Los Angeles, CA 90089, USA

A R T I C L E    I N F O

Article history:
Available online 19 November 2010

Keywords:
Macromolecular assemblies
Visual proteomics
Cryo electron tomography
Cryo electron microscopy
Density fitting
Brownian dynamics simulations
Reaction-diffusion modeling

A B S T R A C T

To increase our current understanding of cellular processes, such as cell signaling and division, knowledge is needed about the spatial and temporal organization of the proteome at different organizational levels. These levels cover a wide range of length and time scales: from the atomic structures of macromolecules for inferring their molecular function, to the quantitative description of their abundance, and spatial distribution in the cell. Emerging experimental technologies are greatly increasing the availability of such spatial information on the molecular organization in living cells. This review addresses three fields that have significantly contributed to our understanding of the proteome’s spatial and temporal organization: first, methods for the structure determination of individual macromolecular assemblies, specifically the fitting of atomic structures into density maps generated from electron microscopy techniques; second, research that visualizes the spatial distributions of these complexes within the cellular context using cryo electron tomography techniques combined with computational image processing; and third, methods for the spatial modeling of the dynamic organization of the proteome, specifically those methods for simulating reaction and diffusion of proteins and complexes in crowded intracellular fluids. The long-term goal is to integrate the varied data about a proteome’s organization into a spatially explicit, predictive model of cellular processes.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

A vast amount of information on a cell’s proteome is being generated by a variety of experimental methods, revealing multiple levels of spatial organization in a cell (Fig. 1) Robinson et al., 2007; Yus et al., 2009; Kühner et al., 2009; Kiel and Serrano, 2009; Collins et al., 2007; Lippincott-Schwartz and Manley, 2009. The molecular structures of individual macromolecules and their assemblies are routinely determined by X-ray crystallography, NMR spectroscopy and electron microscopy (EM). These structures are necessary to determine the molecular functions of macromolecules and their complexes, and provide insights into their evolution (Robinson et al., 2007; Sali et al., 2003). Insights into the proteome organization at a cellular level can be determined by fluorescence imaging (Betzig et al., 2006; Manley et al., 2008) and cryo electron tomography (cryoET) (Leis et al., 2009; Li and Jensen, 2009; Baumeister, 2002). CryoET can be applied to a polymorphic biological specimen in its near native state, and can not only determine a cell’s ultra-structure (i.e., the membrane bound compartmentalization in the cell) (Leis et al., 2009; Briegel et al., 2009; Morris and Jensen, 2008), but also can increasingly produce 3D snapshots of the distributions of large complexes inside the cell (Kühner et al., 2009; Medalia et al., 2002; Beck et al., 2009).

At lower levels of organization, new genomics and proteomic technologies are documenting what genes are transcribed and when (Morris and Jensen, 2008; Yus et al., 2009; Güell et al., 2009). Quantitative mass spectrometry can provide information about which types of proteins are active at a single point in time, and can also reveal the relative and absolute abundance of these proteins and their assemblies in cells and organelles (Kühner et al., 2009; Malmström et al., 2009; Ranish et al., 2003; Aebersold and Mann, 2003). It is also possible to obtain quantitative information about the association rates of reversible protein interactions and the rates of enzymatic reactions in many process pathways (Kinkhabwala and Bastiaens, 2010; Maeder et al., 2007; Dehmelt and Bastiaens, 2010).

A compelling long-term goal of biology is to combine all available spatial and quantitative information about a cell’s constituent parts into a predictive spatially explicit model of cell processes, metabolism and behavior (Morris and Jensen, 2008; Kinkhabwala and Bastiaens, 2010; Lemerle et al., 2005; Ridgway et al., 2006). Ideally, such models would include atomic details relevant to the molecular function of macromolecular assemblies as well as information about the cell’s ultra-structure and the higher-order spatial organization of these assemblies at a cellular level. Both scales are
necessary to understand key biochemical processes. For instance, spatial gradients, irregularities and discontinuities of macromolecular distributions in a cell are known to play an active role in biological processes such as cell division, genome segregation, gene regulation, cell morphogenesis, and shape maintenance. Many biological processes related to cell signaling, cell cycle, and protein transport are also modulated by the precise spatial organization of molecules in space and time (Dehmelt and Bastiaens, 2010; Gavin et al., 2006; Scott and Pawson, 2009; Bork and Serrano, 2005). For instance, in the so-called “push–pull” signaling networks, variations in the spatial distribution of two antagonistic enzymes that covalently modify a substrate can either dramatically weaken or amplify the corresponding signal transduction (Morelli and ten Wolde, 2008). Two other examples are the chemotaxis pathway in E. coli (Lipkow et al., 2005; Bray, 2002), and the mitogen-activated protein kinase (MAPK) cascade processes (Maeder et al., 2007; Dehmelt and Bastiaens, 2010). Even the simplest cells are quite heterogeneous in their cellular proteome organization (Shapiro et al., 2009), but so far relatively little is known about the organization of the cellular proteome at molecular level and how this organization dynamically changes over time.

For a better understanding of cellular processes, it is necessary to expand the scope of structural biology from the molecular functions of the key players in cell processes. These structures provide also a starting point for detecting their localization in the cellular context. Second, cryoET techniques that are used for characterizing the spatial distributions of these assemblies within the cell using computational image processing; and third, spatial modeling of the dynamic organization of the proteome, in particular particle-based reaction-diffusion techniques that simulate the trajectories of these macromolecules and their assemblies over time during reaction processes in their cellular context.

2. Modeling structures of macromolecular assemblies by cryoEM density fitting

High-resolution structures of macromolecular assemblies are key for understanding biological processes at the molecular level (Alber et al., 2008). They can serve as templates in the identification of complexes in whole-cell tomograms (Section 3) and are useful to predict realistic association rates (Schlosshauer and Baker, 2004) and diffusion properties in particle-based simulations of cellular mixtures (Section 4). Although X-ray crystallography has provided spectacular insights into the structure and function of such assemblies, (e.g., the ribosome (Rodnina and Wintermeyer, 2009) and the exosome (Lorentzen et al., 2008)), most of the structural data on abundant large and stable assemblies has been derived in recent years by EM methods (Chiu et al., 2006; Frank, 2009). EM produces 3D density maps of large assemblies typically at intermediate to low levels of resolution (~5–30 Å). The number of density maps in the public domain is growing rapidly, with over 360 structures in the EM databank alone (Henrick et al., 2003). Recent advances in cryoEM data acquisition and image processing...
have made it possible to determine a rapidly increasing number of structures at sub-nanometer resolution, allowing in many cases the identification of secondary structure elements and even the tracing of the backbone (Chiu et al., 2005; Zhou, 2008; Lindert et al., 2009; Stahlberg and Walz, 2008). A number of structures were even solved at near-atomic resolutions (~3.8–4.6 Å), where many detailed structural features such as turns and deep grooves of helices, strand separation in beta-sheets, densities for loops and bulky amino acids side chains could be resolved (Zhou, 2008). Such structures have so far been obtained in highly symmetrical objects such as the bacterial flagellar filament (Yoneykura et al., 2003), various viruses (the cytoplasmic polyhedrosis virus (Yu et al., 2008), rotavirus inner capsid particle (Zhang et al., 2008), the Epsilon 15 bacteriophage (Jiang et al., 2008), the tobacco mosaic virus (Clare and Orlova, 2010), and the GroEL chaperonin (Ludtke et al., 2008). For most cryoEM maps, however, the level of resolution is still not sufficient to directly determine the structure at atomic detail. Indeed, the number of medium- to low-resolution density maps is increasing exponentially while the increase of high-resolution maps grows at a smaller rate. At the intermediate- to low-resolution range, pseudo-atomic models can be obtained by fitting atomic structures and models of individual components (such as whole proteins, domains and nucleic acid chains) into the EM density maps (Stahlberg and Walz, 2008; Orlova and Saibil, 2004; Rossmann et al., 2005).

In this section we review computational methods for rigid fitting, flexible fitting and multiple-component fitting into EM density maps of macromolecular assemblies. We give examples of successful applications of computational methods that produced pseudo-atomic models of macromolecular assemblies.

2.1. Rigid fitting

Rigid fitting methods have been very successful in providing many pseudo-atomic models of macromolecular assemblies (Table 1) (Rossmann et al., 2005; Wriggers and Chacon, 2001; Fabiola and Chapman, 2005; Topf and Sali, 2005). In most of these methods, the quality-of-fit measure for the placement of a probe structure in a density map is the cross-correlation coefficient (or its variants), calculated between the EM density map and the density of the probe structure (computed by convolving the atomic structure using a point-spread function). To find the highest cross-correlation some of these methods (e.g. COLORES (Chacon and Wriggers, 2002)) perform an exhaustive search over all possible probe map positions and orientations relative to the reference EM map. Since performing such a search is computationally intensive, they usually increase the sampling efficiency by applying a Fast Fourier Transform (FFT)-accelerated translational search based on the convolution theorem (Katchalski-Katzir et al., 1992). Sampling efficiency can also be accelerated by a fast rotational search based on spherical harmonics (Kovacs et al., 2003) (e.g., as implemented in ADP_EM (Garzon et al., 2007)). In contrast to the exhaustive search, some programs can perform a heuristic search, for example by using a Monte Carlo-based search algorithm (Wu et al., 2003; Topf et al., 2005).

Other methods increase the computational efficiency by reducing the complexity of density maps to a small set of feature points (so-called code-book vectors) whose positions best reproduce the density map’s gross features, such as its shape and mass distribution (Table 1). The optimal positions of feature points can be determined by the vector quantization (VQ) technique (Martinetz et al., 1993). The fitting problem then effectively reduces into a common point-set matching problem. This matching has been implemented in the STUS package by an exhaustive search method for single-molecule fitting (Wriggers et al., 1999) and a heuristic anchor-point registration method for component fitting into an assembly map (Birmanns and Wriggers, 2007). The latter method uses a hierarchical alignment of the point sets and reduces the search-space complexity by an integrated tree-pruning technique.

There are also a number of programs that allow a local search (e.g., Mod-EM (Topf et al., 2005) and Chimera/Fit_in_Map (Goddard et al., 2007)), that is, maximizing the quality-of-fit measure by searching only in the immediate neighborhood of the initial position of the probe structure. This option is rather useful because often searching for a correct fit of a component in the entire assembly map may lead to an incorrect fit. If knowledge about the approximate position of the component in the map exists (e.g., based on the position of other components or labeling experiments) local fitting may help avoiding this problem. Another way to solve this problem is to use simultaneous multiple-component fitting (see below).

2.2. Flexible fitting

A common problem in EM density fitting is that the isolated component structure may be in a different conformation than in the assembly density map. These conformational differences can originate from the varied experimental conditions under which components and assembly structures were determined as well as from distortions due to crystal packing and experimental noise.

Table 1

<table>
<thead>
<tr>
<th>Computational methods for fitting of atomic structures into cryo-EM density maps.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rigid fitting</strong></td>
</tr>
<tr>
<td>Density-based fitting</td>
</tr>
<tr>
<td>COAN (Volkmann and Hanein, 1999), EMFit (Rossmann, 2000), DOCKEM (Roseman, 2000), FoldHunter (Jiang et al., 2001), Situs/Colores (Chacon and Wriggers, 2002) (Wriggers et al., 1999), UROX (Navaza et al., 2002; Siebert and Navaza, 2009), FRM (Kovacs et al., 2003); grid-threading MC (Wu et al., 2003), 3SOM (Ceulemans and Russell, 2004), Mod-EM (Topf et al., 2005), ADP_EM (Garzon et al., 2007), Chimera/Fit_in_Map (Goddard et al., 2007), Point-matching methods</td>
</tr>
<tr>
<td>SITUS Wriggers et al., 1999 (Birmanns and Wriggers, 2007)</td>
</tr>
<tr>
<td><strong>Flexible fitting</strong></td>
</tr>
<tr>
<td>Fitting multiple conformations</td>
</tr>
<tr>
<td>NMA (Tama et al., 2002; Ma, 2005), NORMA (Suhe et al., 2006; Siebert and Navaza, 2009), Moulder-EM (Topf et al., 2006), S-flexfit (Velazquez-Muriel et al., 2006), Rosetta/Foldhunter (Baker et al., 2006), CHOYCE (Rawi et al., 2010)</td>
</tr>
<tr>
<td><strong>Refinement methods</strong></td>
</tr>
<tr>
<td>RSRef Chen et al., 2003, NMFF/Tama et al., 2004, Flex-EM (Topf et al., 2008), MDFF (Trabuco et al., 2008), Rosetta (Dhibajo et al., 2009), YUP/SCX (Tan et al., 2008), FIRST/RODA (Jolley et al., 2008), DireX (Schoder et al., 2007), DDDF (Kovacs et al., 2008), EM-IMO (Zhu et al., 2010), Cross-correlation based MD (Orzechowski and Tama, 2008), SITUS (Wriggers et al., 1999; Birmanns and Wriggers, 2007; Rusu et al., 2008)</td>
</tr>
<tr>
<td><strong>Simultaneous Fitting of Components</strong></td>
</tr>
<tr>
<td>GEMFIT Kawabata, 2008, MultiFit (Lasker et al., 2009), IQP (Zhang et al., 2010)</td>
</tr>
<tr>
<td><strong>Comprehensive data integration</strong></td>
</tr>
<tr>
<td>IMP (Alber et al., 2008, 2007a,b; Förster et al., 2009; Lasker et al., 2010)</td>
</tr>
</tbody>
</table>
In addition, due to sample heterogeneity single-particle cryoEM and image processing often results in a number of maps describing different functional conformations of the entire complex (Spahn and Penczek, 2009). Common conformational differences are shear and hinge movements of domains and secondary structure elements, as well as loop distortions. When a component structure is generated by a structure prediction method (Baker and Sali, 2001) additional errors can be introduced due to misassignment of secondary structure elements to incorrect sequence regions resulting in their shifts in space (Alber et al., 2008; Topf et al., 2008).

The simplest approach to incorporate conformational variability is to divide the probe structure into rigid bodies, such as domains, and fit each of them independently into the map (Wendt et al., 2009). Common conformational differences are shear and hinge movements of domains and secondary structure elements, as well as loop distortions. When a component structure is generated by a structure prediction method (Baker and Sali, 2001) additional errors can be introduced due to misassignment of secondary structure elements to incorrect sequence regions resulting in their shifts in space (Alber et al., 2008; Topf et al., 2008).

The simplest approach to incorporate conformational variability is to divide the probe structure into rigid bodies, such as domains, and fit each of them independently into the map (Wendt et al., 2001; Volkmann et al., 2000). Clearly, this approach is problematic because the mechanical properties of the probe structure are not reflected in the deformation of the structure. A more objective approach is to generate multiple “valid” conformations of the probe structure and select the top ranking conformation based on its fit in the density. This approach, which essentially separates sampling and scoring, is based on the principle that a correlation exists between the quality of a structural model (e.g., in terms of the RMSD from the native conformation) and its cross correlation with the density map (Topf et al., 2005). An alternative to this approach are methods that simultaneously optimize the conformation of the probe and its position and orientation in the cryoEM map (Fabiola and Chapman, 2005). Such fitting methods are similar to crystallographic refinement programs, except that they generally refine groups of atoms held together instead of individual atoms. Common to all the automated methods is the limited sampling of conformational degrees of freedom. Therefore, they are usually applied to components that are first placed into the density map by rigid fitting.

### 2.2.1. Fitting multiple conformations

In order to generate candidate conformations for fitting one can use normal mode analysis (Tama et al., 2002; Ma, 2005; Suhre et al., 2006), comparative protein structure modeling (Moulder-EM Topf et al., 2006 and the CHOYCE webserver (Rawi et al., 2010), ab initio modeling (Rosetta/Foldhunter, Baker et al., 2006), and exploit the structural variability of protein domains within a given superfamily (S-flexfit, Velazquez-Muriel et al., 2006). Fitting of multiple conformations is particularly useful when crystal structures of the components are not available and must be modeled by comparative modeling or by ab initio protein structure prediction methods. In an example of this approach, most eukaryotic ribosomal proteins were modeled using comparative modeling and fitted into the ribosome density map at ~9 Å resolution (Chandramouli et al., 2008; Taylor et al., 2009) (Fig. 2). A similar approach was implemented using de novo protein structure predictions (Rosetta Rohl et al., 2004). Different conformations were generated and assessed by cryoEM density fitting (Foldhunter Jiang et al., 2001), leading to the discovery of a novel fold for the herpesvirus VP26 core domain (Baker et al., 2006).

### 2.2.2. Refinement methods

Real-space refinement methods (e.g., RSRef (Chen et al., 2003), Flex-EM (Topf et al., 2008), MDFF (Trabuco et al., 2008), Rosetta (Di-Maio et al., 2009), and EM-IMO (Zhu et al. 2010)) use the density to guide the conformational changes in the probe structure. Generally these methods optimize the fit of the probe structure in the density while maintaining its mechanical properties. In Flex-EM for example, the scoring function includes a cross-correlation term as well as stereochemical and non-bonded interaction terms (Topf et al., 2008). A heuristic optimization that relies on a Monte Carlo search, a conjugate-gradients minimization, and simulated annealing molecular dynamics (MD) is applied to a series of subdivisions of the structure into progressively smaller rigid bodies. The method has been used to refine a number of crystal structures and homology models in cryoEM maps, among them the structure of the elongation factor EF4 in the ~11 Å resolution cryoEM map of EF4 bound to the E. coli ribosome (Fig. 3) (Connell et al., 2008) and the structure of the apoptosis-procaspase-9 CARD complex at 9.5 Å (Yuan et al. 2010). In YUP.SCX (Tan et al., 2008) intra-molecular forces are approximated by a Gaussian Network Model (GNM) and the optimization protocol uses simulated annealing MD. Flex-EM and YUP.SCX were applied to refine pseudo-atomic models of proteins and RNA, respectively, using the density of an eukaryotic ribosome at ~9 Å resolution (Taylor et al., 2009) (Fig. 2). In MDFF (Trabuco et al., 2008), external forces proportional to the gradient of the density map are combined with an MD protocol to optimize the conformations of the component structures. This method was recently applied to studies on the mechanisms of protein synthesis by the ribosome (Seidelt et al., 2009) and the dynamics of protein translocation (Becker et al., 2009).

---

**Fig. 2.** Pseudo-atomic models of the yeast 60S (A) and 40S (B) ribosomal subunits determined by combining cryoEM density fitting (at ~9 Å resolution) with RNA and protein homology modeling. YUP.SCX (Tan et al., 2008) and Flex-EM (Topf et al., 2008) were applied to refine the models. Ribosomal RNA is represented as tubes. 5S and 5.8S rRNA are shown as magenta and green tubes, respectively. Ribosomal proteins, with known homology and placement, are shown as orange (A) and pink (B) cartoons. Density accounting for proteins that are unique to eukaryotes with no structural homolog available or that have not been localized within the context of the ribosome is represented in yellow (A) and slate (B). The boundary of each cryoEM map is shown as a gray mesh. (Figure is reproduced from Ref. (Taylor et al., 2009)).
In many refinement methods the conformation is optimized for groups of atoms, either by defining them as rigid bodies (Topf et al., 2008) or by applying restraints on secondary structure elements (Trabuco et al., 2008). Rigid bodies can be selected manually or by automated methods, such as those based on graph theory (e.g., FIRST/FRODA (Jolley et al., 2008)). However, the definition of rigid bodies can limit the conformational degrees of freedom of the structure: if the number of rigid bodies is too small the optimization may not reach the global minimum because a more detailed modification of the conformation is needed. If an all-atom representation is chosen the computational efficiency is largely reduced and the system is likely to get trapped in local minima. A way to avoid this problem can be the use of coarse-grained (CG) or reduced structure representation by considering only Cα atoms (e.g., as in DireX which combines an elastic network model with random walk displacements and distance restraints (Schroder et al., 2007)) or by representing the side-chain of a residue with a pseudo-atom (e.g., as in Damped-Dynamics Flexible Fitting - DDFF (Kovacs et al., 2008)). A recent method applied a Go-model with cross-correlation biased MD without imposing any restraints on the secondary structure elements (Grubisc et al., 2010). A reduced representation can also be used in conjunction with interpolation methods, as an alternative to the constrained MD (Rusu et al., 2008). Finally, conformational changes in a structure can also be guided by normal mode analysis (Tama et al., 2002; Ma, 2005; Suhre et al., 2006). In NMFF for example, a linear combination of low-frequency normal modes is used to iteratively deform the structure to conform to the low-resolution density map (Tama et al., 2004). One of the advantages of this method is that it allows for large conformational changes as observed in the anthrax complex (Tama et al., 2006) or GroEL (Falke et al., 2005). However, NMA is often limited in describing smaller-scale conformational changes (Tama and Sanejouand, 2001) that are often reflected in intermediate-resolution cryoEM density map.

2.3. Simultaneous fitting of multiple components

A considerable challenge is the fitting of multiple components into the density map of an assembly if no a priori knowledge about the location of the components is available. Sequential fitting of components often fails when they cannot be unambiguously placed in the density map as is often the case for maps of intermediate-to low-resolution and for assemblies with a large number of components. In such cases, all the components must be fitted simultaneously into the map to identify the global optimum of the quality-of-fit measure. However, simultaneous fitting is difficult as the large search space makes an exhaustive search protocol (that uniformly samples over all degrees of freedom) computationally unfeasible.

One method (MultiFit (Lasker et al., 2009)) uses discrete sampling in combination with an inference optimizer and expands the quality-of-fit measure by additional information such as shape complementarities between interacting components and the protrusion of components from the map envelope. Other methods, such as GMFIT (Kawabata, 2008) and IQP (Zhang et al., 2010), simplify the search problem by reducing the complexity of the assembly density map and structures (Wriggers et al., 1999; Birmanns and Wriggers, 2007). In GMFIT the initial density distribution of the assembly and the components are approximated by a small set of Gaussian functions to efficiently use gradient-based optimization methods for the structural optimization of component orientations (GMFIT). In the IQP method integer quadratic programming
is used for matching two point sets that represent the reduced density distribution of several components and the assembly. IQP is able to match multiple component point sets simultaneously while considering both information about the geometric architecture of the point distributions as well as the consistency of the density map in the immediate neighborhood of the points. Moreover, the method allows straightforward integration of additional information about the assembly. Reducing the complexity of the density information to point sets is accompanied by an inevitably loss in accuracy in the fitting process. To overcome this challenge, typically a large number of independent point set matches must be performed, which generates an ensemble of candidate solutions. This ensemble is then assessed using an independent scoring function that measures the quality-of-fit between the component structures and the assembly map using the cross-correlation function. Finally, the best scoring structures are refined to locally optimize the fit between assembly and component density maps.

Although progress has been achieved, both simultaneous and flexible fitting of assembly components are still a challenge, in particular for cryoEM maps at resolution worse than ~10 Å (the majority of maps). There is currently a need to develop faster scoring functions that can help discriminate between good and bad fits as well as optimization methods that can explore both conformational and configurational space more efficiently.

2.4. Comprehensive data integration

For the fitting of low-resolution density maps the sampling can be improved if additional information is available about the assembly structure, for instance about protein interactions, sub-complex composition, or shape information from mass spectrometry and SAXS. Recently, a systematic computational framework was introduced (IMP, integrative modeling platform) to determine the structures of large macromolecular assemblies by comprehensive integration of diverse data sources (Alber et al., 2008, 2007a,b; Förster et al., 2009; Lasker et al., 2010). The method can integrate structural information that can vary greatly in terms of their resolution and precision, for instance, data from X-ray crystallography, NMR spectroscopy, electron microscopy, chemical cross-linking, affinity purification, yeast two-hybrid experiments, and computational docking (Alber et al., 2007a; Förster et al., 2009; Lasker et al., 2010). The approach attempts to compute structures that are consistent with all available information about the assembly regarding the composition and structure of the complex. The structure determination process is formulated as an optimization problem where structures that are consistent with all input information are found by minimizing a scoring function. To comprehensively sample the solution space, an ensemble is generated that contains independently calculated structures, each satisfying all of the input restraints.

By applying this framework, the structure of the 450-protein containing nuclear pore complex (NPC) was determined based on the integration of data from seven independent experimental and theoretical sources (Alber et al., 2007a). The averaged structural NPC model has led to functional insights and evolutionary understanding of the yeast NPC (Alber et al., 2007b). Another study determined the pseudo-atomic model AAA-ATPase20S core particle sub-complex of the 26S proteasome (Nickell et al., 2009; Förster et al., 2009). The structure was determined based on a cryoEM map of the 26S proteasome, structures of homologs, and physical protein–protein interaction data.

3. Determining the spatial organization of the cellular proteome

High-resolution structures of macromolecular complexes provide major insights in their molecular mechanism. But to fully understand their function in the context of cellular processes, knowledge of their spatial distribution within living cells is crucial. We currently only begin to understand that the spatial organization of the cell goes far beyond membrane confined compartments and solely diffusion controlled reactions. Cooperating functional modules can function more effectively in local proximity and their reactive sites might be specifically orientated towards each other. For example, recent work demonstrated that certain mRNAs can be locally translated (Lin and Holt, 2008) and that polysomal ribosomes specifically orient along the mRNA trace (Brandt et al., 2009). Such polysomal ribosome arrangements occur in protrusions of Glia cells (Brandt et al., 2010), which might form subcompartments that could allow chaperones to easily access the emerging polypeptide chain and effectively shielding it from the crowded environment until folding is complete. These findings illustrate the need for techniques that reveal the spatial position and orientation of macromolecules at sufficient resolution on a proteome-wide scale.

Classical proteomics approaches determine the protein content of cells, however, any spatial information is lost and cell specific properties are averaged over the population of lysed cells. Recent advances in technology will allow a much better identification and localization of these assemblies in the cell. Among these methods are (i) high-resolution fluorescence imaging, which can provide spatial and dynamic information about the localization, abundance and diffusion rates of macromolecules and their assemblies (Werner et al., 2009) and (ii) cryoET, which not only can provide insights into the ultra-structure of cells but increasingly even into the cellular distribution of assemblies (Beck et al., 2009; Ben-Harush et al., 2009; Förster et al., 2010; Grünewald, 2009). With these methods it may be possible to gain information about the cellular environment of macromolecular assemblies, and gain insights into the positions of all their instances in a cell, during different time points in its lifetime. They also promise to reveal differences in the distribution of macromolecules between individual cells and their contribution to the behavior of the entire population. Such information will help bridging the knowledge about the molecular function of macromolecules derived from their atomic structure with their cellular function at the cellular systems level. In the following section we focus on methods and applications of cryoET.

3.1. Visual proteomics using cryo electron tomography

CryoET provides 3D snapshots of cells under close-to-live conditions (Figs. 4 and 5). However, assigning the protein identity to the observed shapes and distribution of complexes is not a trivial task due to a number of technical challenges: a tomogram typically has low signal-to-noise ratio, non-isotropic resolution, and straightforward in vivo labeling techniques, analogous to GFP-tags used in fluorescence imaging, are not available for cryoEM (Leis et al., 2009; Morris and Jensen, 2008; Nickell et al., 2006; Lucic et al., 2005). For the proportion of protein assemblies whose structures are known, visual proteomics methods have been developed that attempt to generate molecular atlases containing the position and angular orientation of protein complexes in cryo electron tomograms of cells by template matching (or pattern recognition) (Baumeister, 2002; Nickell et al., 2006; Förster et al., 2010; Frangakis et al., 2002; Böhm et al., 2000). At first a template library is built that contains the reference structures of the protein complexes to be localized. Subsequently, an exhaustive search is performed that scans the tomographic volumes for matches with the structural templates by calculating a cross-correlation score. Finally, the observed distribution of the cross-correlation score is translated into a position list using peak extraction and statistical methods. In 2000, the first proof-of-concept for template matching
Visual proteomics of the human pathogen *Leptospira interrogans*. (A) Score distributions obtained from artificial tomograms of the observed (blue), true positive (green) false positive hits (red). The arrowhead marks the threshold that accounts for a specificity of 40%. (B) Centered slices through an artificial tomogram shown perpendicular (left) and parallel (right) to the electron optical axis, as well as before (top) and after (bottom) simulating the image formation process. Numbers account for protein complexes as described below. (C) Centered slice through a tomographic reconstruction, oriented perpendicular to the electron optical axis. (D) The templates assigned into this volume are shown surface rendered. 70S ribosomes (1) are shown in dark brown, 2 RNA-polymerase II (2) in green, ATP-synthase (3) in bright brown, GroEL in bright blue, GroEL-ES (4) in dark blue, Hsp (5) in red, cytoplasmic membrane in transparent blue, and the cell wall in transparent brown. (Figure is adapted from Reference (Beck et al., 2009)).

Visual proteomics approaches are so far dependent on the availability of template structures, or, with other words, are biased by their own template libraries. In order to generate libraries of template structures that best account for the signal observed in a cryo electron tomogram, the imaging process has to be simulated as realistically as possible. First, the imaged quantity needs to be calculated from the coordinates and identities of the atoms specified in the PDB. Subsequently the density is convoluted with the contrast transfer function, which describes the imaging in the transmission electron microscopy in linear approximation (Forster et al., 2010; Lucic et al., 2005). Beyond these technicalities, there are a few biological aspects to be taken into account when template libraries are being assembled. Can the template sufficiently account for the structural state of the targeted molecule within the cell? A recent structural-proteomic survey of the largest and most abundant protein complexes in *Desulfovibrio vulgaris* revealed a high structural diversity between different microorganisms (Han et al., 2009). The selection of reference structures solved from other species has therefore to be done with caution. Furthermore, the protein complex might exist in diverse oligomeric states in the cell that might differ from the species observed during structure determination. For this reason, visual proteomics approaches so far were limited to highly stable and conserved targets. What is the abundance of the targeted protein complexes in the cell and to what extent does the template library account for all large protein complexes that exist in the cell? This is an important question as well, because high abundant protein complexes of similar structure might compete for assignments and thereby have a negative effect on the performance. Targeted and directed mass spectrometric measurements can provide the required auxiliary information about protein abundances on a proteome-wide scale (Malmstrom et al., 2009).

### 3.1.2. Assessment of template matching

Once the template library is built, matches are identified by calculating a cross-correlation score between reference structures and the subvolume of a tomogram. Förster implemented a local, constrained correlation function for template matching to account for missing wedge effects (due to incomplete tomogram sampling) (Forster, 2005; Förster et al., 2005; Roseman, 2004; Ortiz et al., 2006). Recently, a scoring function was introduced that not only evaluates the cross-correlation value for the template, but also the cross-correlation values of competing templates and decays at the same position in the tomogram (Beck et al., 2009). This scoring scheme outperformed the classical workflow particularly for competing protein complexes of similar structural signature. For visualization, the template is rotated and positioned at the determined coordinates and the corresponding angles that lead to the maximal cross-correlation values. The major challenge thereby is to discriminate true positive from false positive detections (Fig. 4A). The performance not only depends on tomogram-specific parameters such as acquisition settings, specimen thickness, and molecular crowding, but also target-specific parameters, such as molecular weights, cellular abundance, and cellular abundance of protein assemblies competing for assignments. Therefore, the performance can vary for each template contained in the library. The false positive discovery rates of all matches have to be estimated using a statistical model (Beck et al., 2009; Keller et al., 2002). The performance of template matching can also be assessed based on a priori knowledge about the spatial distribution of the template structures, such as the cellular localization or orientation of protein assemblies, which can provide clues about the expected false positive discovery rates. For example, some membrane associated protein complexes exhibit specific positioning and orientation relative to the membrane (Ortiz et al., 2006). Also matching with a mirrored template of ‘non-native’ handedness can provide complementary information about the distribution of cross-correlation values. In addition, simulated tomograms can serve as test cases to estimate the performances. Because the position and orientation of all protein complexes in the artificial tomograms are known, the false positive rate can be calculated for specific template complexes under the specific experimental conditions (Fig. 4B).

### 3.2. Applications

In a recent study, namely the visual proteomics project for the pathogen *Leptospira interrogans* (Beck et al., 2009; Förster et al., 2010; Grünewald, 2009; Malmstrom et al., 2009), targeted proteomics experiments detecting the identity and concentrations of cellular proteins were combined with cryoET-based template matching to detect the spatial localizations of a set of protein complexes (Fig. 4). To identify template structures that might be suitable for the template matching, 26 candidate assemblies of a
certain minimal size (>250 kDa) and a certain sequence conservation were retrieved from structural databases (Beck et al., 2009; Förster et al., 2010). Targeted proteomics on a proteome-wide scale determined that the cellular protein abundance in *L. interrogans* ranged over 3.5 orders of magnitude (Malmstrom et al., 2009). Only ten other protein complexes of sufficient size for template matching were found with abundances of at least 100 copies per cell. Tomograms of *L. interrogans* cells were acquired and subjected to template matching and scoring (Fig. 4C and D). The local concentration of the targeted protein complexes varied within and across data sets: the cells displayed an average ribosome concentration of \( \frac{1}{24} 20 \text{M} \) in the cytoplasm, but the local concentration ranged from 5 to 30 \( \frac{1}{24} 100 \text{M} \). The local fluctuations in case of total GroEL together with GroEL-ES were larger and ranged from 8 to 100 \( \frac{1}{24} 0.5–6.5 \text{mg/ml} \). This study showed that: (i) background noise reduces the performance depending on the MTF of the particular CCD camera; (ii) the missing wedge can introduce angular bias to the discovery rate of templates with anisotropic signal content (such as the elongated ATP-synthase); (iii) the specificity increases with the template abundance; and (iv) decreases with the degree of molecular crowding (Beck et al., 2009; Malmstrom et al., 2009). The last two effects are generally more pronounced for smaller than for larger templates. The conclusion was that the specificity achieved for high abundant Megadalton complexes is satisfactory. However, true-positive discovery rates higher than 50% are difficult to achieve when protein complexes of smaller molecular weights are targeted. For complexes of very low cellular abundance it is quite a challenge to obtain robust statistical models (Beck et al., 2009; Malmstrom et al., 2009).

Recently, the proteome organization of the small model bacteria *M. pneumoniae* was analyzed, revealing its proteome composition and organization, as well as its transcriptional and metabolic regulation (Yus et al., 2009; Kühner et al., 2009; Güell et al., 2009). Proteomics-based TAP-mass spectrometry data were complemented with structural modeling and cryoET to provide insights into the structural anatomy of the model bacteria (Kühner et al., 2009). The study identified more than 170 protein complexes. Four of these protein complexes whose structures have been known were localized in the tomogram of an intact *M. pneumoniae* cell (Fig. 5). This ratio of detected to structurally-characterized protein complexes demonstrates current throughput limitations of cryoEM arising from: (i) the stability of protein complexes during the cryo preparation; (ii) their orientation on the EM-grid and; (iii) the fact that the data analysis by image processing is time consuming and requires intensive user interference. Nevertheless, the study demonstrates that mapping of macromolecular structures into entire-cell tomograms is a powerful strategy when combined with unbiased large-scale complex purification techniques. The template library in this study was in part built from single-particle EM structures of *M. pneumoniae* protein complexes. This considerable effort overcomes the difficulty of structural conservation across species discussed above.

CryoET combined with image averaging provides unique opportunities for structure determination of membrane-bound macromolecular assemblies in intact cells (Bartesaghi and...
Subramaniam, 2009; Förster and Hegerl, 2007). For instance, a recent study investigated receptor clustering in bacterial chemotaxis and found a universal architecture in a wide range of bacteria (Briegel et al., 2009, 2008). In addition, cryoET can provide snapshots of individual time points along a process pathway, as has been demonstrated for cytoskeletal-driven processes that are involved in cell movements (Ben-Harush et al., 2010) and membrane fusion in herpes simplex virus 1 entry (Maurer et al., 2008).

3.3. Outlook

A number of limitations to visual proteomics might be overcome by further technical developments. Currently the most critical limitation in cryoET is the moderate signal-to-noise ratio. Since the dosage of electrons that can be applied to a biological specimen ultimately limits the resolution of cryoET, visual proteomics currently remains restricted to complexes of a certain minimal molecular weight that can be targeted by template matching. However, the signal in visual proteomics primarily arises from the contrast between the targeted protein complex and the surrounding solvent. Therefore, improving contrast, e.g. through phase plates (Murata et al. 2010), is likely to improve the performance in the future. Furthermore, specimen thinning techniques, e.g. through focused ion beams (Rigort et al., 2010), will be crucial for improving SNR and imaging larger cell types, such as eukaryotic cells, which exceed 500 nm in diameter; thus, application of visual proteomics to eukaryotic cells depends on progress in this field. Finally, direct electron detection systems promise to largely improve SNR of cameras (McMullan et al., 2009), which will be highly beneficial for techniques that target pleomorphic structures, such as cryoET, in which the SNR cannot be improved by imaging the multiple copies of the same specimen.

However, some limitations will remain: cryoET provides static snapshots of single cells without temporal resolution and the spatial resolution of live cell imaging techniques is comparably low. However, cryoET information can be complemented by computational modeling of the proteome diffusion and reaction processes. Such combined approaches hold great potential to reveal insights into the dynamic spatial organization of the proteome and the biological processes in cells.

4. Modeling the dynamic proteome organization

Novel experimental methods have greatly increased the availability of information on the molecular organization of living cells. As described in the previous sections information is increasingly available about the membrane confined compartments of cells and the identity, abundance and spatial distributions of the macromolecular components in the cell. There is a pressing need to integrate these varied data into spatially explicit, predictive models of the cell. Molecular components in the cell geometries into a network model of reaction processes. As a consequence, spatial gradients and localized interactions such as cell division and nucleo-cytoplasmic transport, are either spatially constrained or segregated (Terry et al., 2007). These activities actively exploit spatial gradients in the macromolecular distributions. Spatial network models, such as Virtual Cell (Schaff et al., 1997; Slepchenko et al., 2003), MesORD (Hattne et al., 2005), and SmartCell (Andr et al., 2004) incorporate such spatial effects by partitioning the cell volume into microdomains (Lemerle et al., 2005; Ridgway et al., 2006). Virtual Cell (Schaff et al., 1997; Slepchenko et al., 2003), can incorporate experimentally determined cell geometries into a network model of reaction processes.

When biomolecules are present in relatively low copy numbers, their local concentrations can fluctuate widely, which can cause stochastic effects in reaction processes. The effective behavior of such molecules may be very different from their behavior under a constant distribution; furthermore, in such cases the law of mass action for reaction kinetics no longer applies. It has been shown that such stochastic particle behavior can significantly influence gene regulation (Raser and O’Shea, 2005; Cai et al., 2006), signal transduction (Kollmann et al., 2005) and many other processes (Ridgway et al., 2006). To model stochastic behavior several network models (including SmartCell (Andr et al., 2004) and MesORD (Hattne et al., 2005)) replace the law of mass action with a chemical master equation (Ridgway et al., 2006; Turner et al., 2004). Master equations are also simulated by the Stochastic Simulation Algorithm (SSA) (Gillespie, 1977), and its variants such as the tau-leaping method (Rathinam et al., 2003), which can account for low copy numbers in cellular network simulations.

4.2. Particle-based reaction–diffusion simulations

Particle-based simulations naturally incorporate the concepts of space, crowding and stochasticity, so often provide a more realistic treatment than network models. A protein or other reactants are often treated as single particles. Every particle in the reaction system is treated explicitly, and the particle positions are sampled at discrete time intervals. To simulate particle diffusion, these simulations use a method known as Brownian dynamics (BD) (Ridgway et al., 2006; Ermak and McCammon, 1978): the net force experienced by a particle contains a random element in addition to contributions from interactions with other particles. The random element is an explicit approximation to the statistical properties of Brownian forces, due to the effects of collisions with solvent molecules, which are not explicitly modeled. More specifically, the particles are displaced from their position at each time interval by a random vector whose norm is chosen from a probability distribution function that is a solution to the Einstein diffusion equation. As a consequence, spatial gradients and localized interactions naturally occur. A number of reaction–diffusion algorithms incorporating Brownian dynamics have been developed (Morelli and ten Wolde, 2008; Ridgway et al., 2008; Boulanne et al., 2008; Andrews and Bray, 2004; Sun and Weinstein, 2007). Interactions and reactions occur upon contact between particles according to specific probabilities, which are chosen to reproduce the correct reaction kinetics (Morelli and ten Wolde, 2008; Ridgway et al., 2008; Boulanne et al., 2008; Andrews and Bray, 2004). The particle-based simulation scheme, described above, has been implemented...
by several packages, including GridCell (Boulianne et al., 2008), MCell (Coggan et al., 2005), Smoldyn (Andrews and Bray, 2004), and the molecular simulators described by Ridgway et al. (2008) and Morelli and ten Wolde (2008). In GridCell, particles move only to occupy vertices of cubic grid structures. In MCell and Smoldyn the molecules move freely and are treated as point particles without exclusion volume. Thus the latter two methods do not explicitly resolve particle collisions. In Smoldyn, reactions are accepted if two particles pass within a certain distance chosen such that the overall reaction rate is reproduced correctly. As the particles lack volume, macromolecular crowding in the cell interior is incorporated indirectly by adding inert stationary, impenetrable blocks to the cytoplasm (Lipkow et al., 2005; Andrews and Bray, 2004). Smoldyn can be applied with relatively long time steps, allowing simulations at biologically relevant time scales. Smoldyn was used to study the effects of cellular architecture on signal transduction in E. coli chemotaxis (Lipkow et al., 2005; Lipkow, 2006) (Fig. 6A). Simulations were performed for the diffusion of CheYp from the cluster of receptors to the flagellar motors, first under control conditions and then in response to attractant and repellent stimuli (Lipkow et al., 2005). Smoldyn has also been used to study the signal amplification in a lattice of coupled protein kinases (Goldman et al., 2009). MCell is based on Monte Carlo simulations, and can incorporate specific shapes and cell features derived from cryoET and other experimental methods. MCell has been used to study presynaptic calcium dynamics and neurotransmitter release at a neuronal synapse (Coggan et al., 2005). Another study simulated macromolecular transport through the nuclear pore complex to interpret the experimentally determined cargo distribution in light of existing models for nuclear import (Beck et al., 2007).

The intracellular space is a highly crowded and non-uniform with an occupied volume of typically up to 50% in eukaryotes and 30–40% in bacteria (Ridgway et al., 2006; Minton, 2006; Ellis and Minton, 2003) (Fig. 6 B and C). In order to accurately simulate crowding effects, a number of reaction–diffusion methods take into account the physical dimensions of proteins and complexes (Ridgway et al., 2008; Sun and Weinstein, 2007). Most of them model the proteins and complexes as hard, elastic spheres (Ridgway et al., 2008; Sun and Weinstein, 2007; Kim and Yethiraj, 2009). Crowding can have several profound effects on the dynamics within the cell, such as anomalous subdiffusion of macromolecules (Golding and Cox, 2006; Weiss et al., 2004). These effects in turn have a direct impact on the thermodynamics and kinetics of biological processes, including protein interactions, diffusion, stability, folding, and biochemical reactions rates (Ellis and Minton, 2003; Schreiber et al., 2008). Crowding slows rates of diffusion-limited reactions (Kim and Yethiraj, 2009; Schreiber et al., 2008; Wieczorek and Zielenkiewicz, 2008), but accelerates reactions with low association rates (i.e., very small reaction probabilities per collision), because it forces colliding particles to remain longer in closer proximity. However, simply giving the particles an excluded volume is not sufficient to accurately reproduce crowding effects.
about the molecular functions of the cellular components as well as quantitative information about their abundance, interactions and spatio-temporal distributions in the confined geometry of the cell. Emerging new experimental technologies are greatly increasing the availability of such spatial information about the molecular organization in living cells. In this review we have addressed three distinct fields that have contributed to our understanding of the proteome’s spatial and temporal organization. CryoEM density fitting methods can provide pseudo-atomic structures of macromolecular assemblies. Such structures provide insights into the molecular functions of assemblies and can also serve as templates for visual proteomics approaches, which can potentially identify their abundance and distributions in the cell and therefore provide a three-dimensional view of the cellular proteomes and interactomes. Although the method is currently limited to small cells and large macromolecular assemblies, a number of limitations to visual proteomics might be overcome by further technical developments, which will eventually allow to expand this exciting technology to larger eukaryotic cells. Finally, ongoing research is reviewed on the dynamic modeling of cellular reaction processes. Over the last few years several particle-based reaction diffusion methods have been developed that accurately reproduce protein diffusion and reactions in highly crowded cellular environments. However, it is an ongoing challenge to extend such methods to cover the wide range of time scales seen in biological processes, while maintaining high accuracy in reproducing the reaction properties. The long-term goal will remain to integrate information derived at the various levels of spatial proteome organization, into a predictive spatial model of cell processes. Such integration will be facilitated by the continued extension of structural biology methods from the molecular to the cellular level.

Acknowledgments

The authors acknowledge financial support from the Human Frontier Science Program (RGY0079/2009-C to F.A. and M.T.), the National Institute of Health (NIH) (1R01GM096089 and 2U54RR022220-06 to F.A.), Alfred P. Sloan Research foundation (to F.A.); F.A. is a Pew Scholar in Biomedical Sciences, supported by the Pew Charitable Trusts, Medical Research Council Career Development Award (G0600084) (to M.T.) and the ‘Special Presidential Prize and Excellent PhD Thesis Award – Scientific Research Foundation of the Chinese Academy of Sciences (to S.Z.). This review is partially based on refs. (Beck et al., 2009; Alber et al., 2008; Förster et al., 2010). The authors also acknowledge M.S. Madhusudhan and Ben Mathiesen for assistance in preparing the manuscript and useful discussions.

References


