Archives of Biochemistry and Biophysics xxx (2016) 1–9

Contents lists available at ScienceDirect

Archives of Biochemistry and Biophysics

journal homepage: www.elsevier.com/locate/yabbi

Computational crystallization

Irem Altan [a], Patrick Charbonneau [a, b], Edward H. Snell [c, d, *]

[a] Department of Chemistry, Duke University, Durham, NC 27708, USA
[b] Department of Physics, Duke University, Durham, NC 27708, USA
[c] Hauptman-Woodward Medical Research Institute, 700 Ellicott St., NY 14203, USA
[d] Department of Structural Biology, SUNY University of Buffalo, 700 Ellicott St., NY 14203, USA

ARTICLE INFO

Article history:
Received 4 December 2015
Received in revised form
22 December 2015
Accepted 7 January 2016
Available online xxx

Keywords:
Crystallization
computational
crystallization data
crystallography
macromolecular

ABSTRACT

Crystallization is a key step in macromolecular structure determination by crystallography. While a robust theoretical treatment of the process is available, due to the complexity of the system, the experimental process is still largely one of trial and error. In this article, efforts in the field are discussed together with a theoretical underpinning using a solubility phase diagram. Prior knowledge has been used to develop tools that computationally predict the crystallization outcome and define mutational approaches that enhance the likelihood of crystallization. For the most part these tools are based on binary outcomes (crystal or no crystal), and the full information contained in an assembly of crystallization screening experiments is lost. The potential of this additional information is illustrated by examples where new biological knowledge can be obtained and where a target can be sub-categorized to predict which class of reagents provides the crystallization driving force. Computational analysis of crystallization requires complete and correctly formatted data. While massive crystallization screening efforts are under way, the data available from many of these studies are sparse. The potential for this data and the steps needed to realize this potential are discussed.

1. Introduction

Macromolecular crystallography is a gateway to the detailed structure of biological molecules, and, to the biological processes in which they are involved. Its power as a scientific tool is recognized in the remarkable number of Nobel prizes that make direct reference to the technique. The main drawback of the approach is that – as its name indicates – it requires a crystal of the macromolecule of interest. There lies the crux of the problem.

The National Institutes of Health Protein Structure Initiative (PSI) targeting fold space (where all outcomes, crystallization and non-crystallization, were tracked) showed that out of ~45 K soluble, purified targets, ~8 K crystallized, of which only ~5 K resulted in a crystal structure. Later data from the PSI-Biology initiative, which focused on targets of compelling biological interest, showed that out of a further ~10 K targets, only ~2 K resulted in a crystal structure [1]. Results from these large datasets mirror those obtained at the large-scale crystallization screening center led by one of us: the Hauptman-Woodward Medical Research Institute High-Throughput Screening (HTS) Laboratory [2]. An analysis of 96 biological macromolecular targets screened against a set of 1536 chemical cocktails gave 277 crystal leads (36 targets produced one or more crystals) from ~150 K experiments, and hence only ~0.2% of the experiments used for screening produced crystals [3,4]. These results thus provide a consistent illustration of the remarkably poor success rate of crystallization screening despite the power of the crystallographic technique itself.

A meta-analysis of large-scale crystallization screening centers reveals that out of ten soluble protein constructs, four are likely to crystallize, of which only one on average will produce a crystal structure [5]. Because the number of structures deposited in the Protein Data bank (PDB) that have been determined by X-ray crystallographic methods is now over 100 K, we infer that tens of millions of experiments have had outcomes other than diffracting crystals. Unfortunately, information on those experiments is not reported in the PDB, and in most cases nor are the conditions that produced a crystal but were not used in the final structure determination. We are thus left with datasets that incompletely capture the effort involved in the crystallization of macromolecules. If that were not bad enough, the data that is available often presents

* Corresponding author. Hauptman-Woodward Medical Research Institute, 700 Ellicott St., NY 14203, USA.
E-mail address: esnell@hwi.buffalo.edu (E.H. Snell).

http://dx.doi.org/10.1016/j.abb.2016.01.004
0003-9861/

multiple difficulties for automated attempts at data mining and prediction. In this article, we first describe the crystallization process, and then show that by utilizing the PDB-deposited structures and sequences along with the information whether or not that sequence produced a crystal, it is possible to provide predictive glimpses of the experimental outcome. We then review attempts at employing detailed screening information in order to provide further insight. We conclude with a discussion of the incentives for improved bookkeeping and with an outlook towards future research in computational crystallization.

2. Crystallization

Crystallization has been described as an "empirical art of rational trial and error guided by past results" [6]. For better and for worse, this still sums up the state of the art of the field. When presented with a biological macromolecular sequence it is currently impossible to know what conditions will result in crystallization or even if crystallization will occur at all. Basic predictive techniques can determine, for instance if the target may be associated with a membrane, providing an initial lead into the appropriate crystallization approaches, but little further guidance is available. The history of crystallization methods and the development of modern crystallization screens have been summarized elsewhere [7]. It is, however, important to note that the development of readily prepared commercial screens make initial crystallization screening experiments easier, but at the risk of standardizing the starting points and introducing a bias on outcome. We will not discuss the details of the available screens or approaches here, but rather sketch the features of these methods that are essential to describing the role and purpose of experimental crystallization data.

The main strategy in macromolecular crystallization is to gradually bring a target solubilized in an appropriate aqueous chemical cocktail (containing at least a precipitating agent and a buffer) to a region of supersaturation until a crystal nucleates and grows. Interestingly, experimental outcomes reveal that there are many landmarks on the screening landscape that are not crystals. Crystals form under solution conditions that fall between those producing clear drops (conditions under which macromolecular–solvent interactions are stronger than interactions between macromolecules) and those producing precipitate (conditions under which the opposite is true). Clear and precipitate are common outcomes but phase separation, skin-formation, and a combination of any of these outcomes can also occur. These outcomes are markers on the solubility landscape that can be used to direct experiments towards a crystal [8]. For the vast majority of reports, however, these outcomes are not captured, restricting significantly the insights that can be gained from failed crystallization experiments into profiling the phase response of a macromolecule to different chemistries.

To understand the potential insights it is useful to review the thermodynamical framework of crystallization screening experiments. First and foremost, macromolecular crystallization is a phase transition. To study the coexistence of the crystal and the solution forms of a macromolecule, one needs to determine the set of conditions at which the chemical potentials of the two phases are equal. The chemical potential determination from first-principles entails knowledge of effective interactions between macromolecules [9–12]. Unfortunately, although the individual forces involved in macromolecular interactions are well understood, there does not exist a simple and fast way to figure out how these forces collectively build up effective macromolecular interactions [13,14].

Both experimental and theoretical studies of crystallization inevitably involve the interpretation of phase diagrams. These recapitulate the conditions under which different phases are thermodynamically stable. For mixtures such as crystallization cocktails, *i.e.* the mixture of chemicals designed to drive crystallization, the phase diagram is multi-dimensional. While not true in general, these multi-dimensional macromolecular phase diagrams can often be projected on the temperature-macromolecular concentration plane. This projection changes the temperature scale, however, as the contents of the cocktail, such as co-solutes, alter the energy scale of macromolecular interactions (while the experimental temperature itself is barely changed). The composition of the final phase can also be complex. A macromolecular crystal, for instance, contains on average more than 50% of the mother liquor [15] and may embed co-solutes in fractions that differ from what is left in the crystallization cocktail.

The arduous task of experimentally determining phase diagrams was carried out for a handful of proteins, including lysozyme, γ-crystallins, insulin, and myoglobin [16–22]. Despite the complexities introduced by the macromolecular structure and the additives in the crystallization cocktail, the phase diagrams of these proteins resemble that of simple liquids [23], in the sense that phases analogous to the crystal, liquid, vapor, and supercritical fluid are distinguishable. The phase diagram of simple liquids (shown in Fig. 1a), however, exhibits a vapor–liquid coexistence line that terminates at a stable critical point. Below this line (in the dark gray area) the system is unstable, and above the critical temperature, it is a supercritical fluid. One key difference is that for proteins the critical point typically lies below the crystal solubility line, meaning that the protein vapor and liquid phases are metastable. For proteins (Fig. 1b), below the critical point, the system has a propensity to aggregate in a disordered, percolating network, *i.e.*, a gel; the gelation probability increasing with distance to the critical point [24,25]. Although such an aggregation is metastable, it can be long-lived, which arrests the crystallization process in the timescale of experiments. The area between the solubility line and the critical point is where crystallization is most likely to occur, and has consequently been dubbed the nucleation zone or crystallization gap [26,27].

Macromolecular phase diagrams are informative of the conditions in which crystallization is most likely to occur. However, their experimental determination is a task more challenging and time-consuming than even the most ambitious crystallization screen. Furthermore, theoretically determining phase diagrams to guide crystallization experiments suffers from the fundamental caveat that the macromolecular structure is required, because effective macromolecular interactions cannot be reliably determined

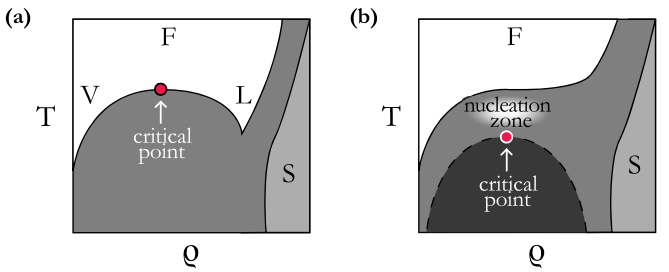

Fig. 1. Phase diagram of (a) a simple liquid and (b) a simple protein in the density–temperature plane (T-ρ). For crystallization cocktails, the projection of the multi-dimensional phase diagram onto the T-ρ plane typically involves rescaling the temperature axis by an aggregate function of the solution conditions. Vapor, liquid, solid, and fluid phases are denoted by V, L, S, and F, respectively. For macromolecules, the critical point [73] lies below the solubility line; for simple liquids, this critical point is located at the top of the (hidden) solubility curve. A supersaturated protein solution is typically found to be most likely to crystallize in the nucleation zone — the region between the solubility line and the critical point.

Please cite this article in press as: I. Altan, et al., Computational crystallization, Archives of Biochemistry and Biophysics (2016), http://dx.doi.org/10.1016/j.abb.2016.01.004