Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Statistical experimental design of protein crystallization screening revisited

Thanh Tam Tran,[a] Isabelle Sorel[b] and Anita Lewit-Bentley[a]*

[a]LURE, CNRS, CEA, Ministère de la Recherche, BP 34, 91898 Orsay CEDEX, France, and [b]IBBMC, Bâtiment 430, University Paris XI, 91405 Orsay, France

Correspondence e-mail: anita.bentley@lure.u-psud.fr

Received 8 March 2004
Accepted 28 June 2004

A statistical experimental design approach was used to prepare a set of solutions for the screening of protein crystallization conditions. This approach is shown to be amenable to quantitative evaluation and therefore to the rational optimization of the screening results. All solutions contain a cryoprotectant, thus eliminating the need for subsequent optimization of crystal freezing conditions.

1. Introduction

The successful crystal structure determination of a biological macromolecule requires the preparation of diffraction-quality crystals. Despite the enormous progress made in the crystal structure determination process itself, as well as some spectacular recent successes concerning difficult problems such as membrane proteins and very large macromolecular assemblies, the underlying crystallization methodology has remained empirical. More recently, structural genomics projects have developed high-throughput techniques that can be applied to all stages of crystal structure determination, including the automation of as many of them as possible. This is particularly true for crystallization experiments, where the sheer number of trials has become such that it can be impossible to prepare and manage them by traditional manual means (Chayen & Saridakis, 2002).

Protein crystallization depends on a number of variables, given by the components of the protein solution (pH, buffer composition, salts, precipitating agents *etc.*) as well as external parameters such as temperature. Screening for the optimal set of conditions is thus a multiparameter problem, which has been addressed by many researchers using different approaches. It is, furthermore, a problem that should be amenable to statistical methods, a fact that was first recognized by C. W. Carter (Carter & Carter, 1979; Carter, 1990), who pioneered the use of statistical experimental design. This approach has since been used by Abergel *et al.* (1991), who have made a software server available on the Internet for the practical design of crystallization screening experiments (Audic *et al.*, 1997). A more recent system accessible on the web has been proposed by Segelke (2001). In both cases, however, the number of components the experimenter can combine is limited.

A parallel development was that of the 'sparse-matrix' approach, *i.e.* sampling a large number of possible crystallization conditions with a limited number of combinations of their ingredients (Jancarik & Kim, 1991). The advantage of this scheme lies in its versatility and ease of use: the 50 solu-

© 2004 International Union of Crystallography
Printed in Denmark – all rights reserved

tions to test were selected on the basis of known and successful crystallization conditions and combined to sample a wide range of precipitant, buffer and salt types. These solutions are today marketed by Hampton Research as a screening kit (Crystal Screen I), together with a second kit of slightly different composition, and similar kits have been devised by other manufacturers.

While the commercial crystallization screening solutions have the advantage of being readily available and convenient for use manually as well as in robotics applications, unless their use gives a clear-cut result the strategies for optimization of the conditions are not at all obvious. This is because of the nature of these kits: each solution is composed of different ingredients with little repetition of individual ingredients within the kit. Thus, it is not clear which of the components of a given solution is responsible for the positive result and all have to be optimized. Furthermore, in the case of Crystal Screens by Hampton Research, it has been shown that the effective pH of the solutions may differ significantly from that announced and can vary with time (Bukrinsky & Navarro-Poulsen, 2001; Wooh *et al.*, 2003).

We have tried to combine a rigorous statistical experimental design with the advantages of sampling a relatively large number of typical components used today in the screening for crystallization conditions. With this approach, we retain the versatility of crystallization screening kits with the capacity for statistical evaluation of the experimental results, thus allowing rational optimization of the initial conditions found. We expect that this method will also be useful in robotics applications.

2. Materials and methods

The experimental designs were prepared and evaluated using the program package *MODDE*6.0 from Umetrics, Sweden. All chemicals used for the preparation of crystallization solutions were of analytical grade purity. The following proteins were used without further purification: bovine serum albumin (BSA), human and canine apotransferrin, lysozyme, phospholipase A_2, triose phosphate isomerase, concanavalin A and ubiquitin from Sigma–Aldrich, glucose isomerase and xylanase from Hampton Research, α-chymotrypsin from Buchs, insulin from Novo, trypsin from Serva and papain from Worthington Biochemicals. ORFs Nos. 69 (YIL020c), 116 (YER067w), 182 (YD2435c) and 190 (YOR357c or Grd19p; Zhou *et al.*, 2003) are among the French Yeast Structural Genomics Intiative targets (http://www.genomics.eu.org). All experiments employed the vapour-diffusion technique. We used a Tecan robotics station for all the crystallization screening experiments, with Greiner 96-well/388-drop plates sealed with a transparent tape. All crystallization solutions were sterile-filtered and kept at 277 K. The optimization experiments were carried out by hand, using hanging-drop vapour diffusion and 24-well Linbro plates. The crystallization trials were observed regularly under a stereo-microscope until no further evolution was seen.

3. Results

3.1. Selection of crystallization solution components

The following types of components were included: buffer (*i.e.* defined pH), 'organic' precipitating agent (*i.e.* PEG, MPD or 2-propanol), salt, divalent cations and additives. We further included a cryoprotecting agent within all of the crystallizing solution formulations, since crystallographers routinely use data collection at liquid-nitrogen temperatures. The search for an appropriate type and concentration of cryoprotectant can be time- and crystal-consuming (Garman, 1999) and could be eliminated if the crystal mother liquor were itself a cryoprotectant solution. The combination of organic and salt precipitants at different concentrations has been shown to be very useful in producing favourable crystallization conditions (Majeed *et al.*, 2003) and is included in our designs. We chose to limit our designs to 48 distinct solutions, a number typically accessible to present-day robotics systems: each experiment is carried out twice to produce a 96-well plate. The limitation to 48 different solutions and five different component types restricted our choice of possible compounds. The following choice was made based on compounds used most frequently in crystallization experiments (see, for example, the macromolecular crystallization database, BMCD; http://wwwbmcd.nist.gov:8080/bmcd/bmcd.html; Page *et al.*, 2003; Wooh *et al.*, 2003): high-molecular-weight PEG (PEG 8000), low-molecular-weight PEG (PEG 400), MPD and 2-propanol (isopropanol) as 'organic' precipitants, ammonium sulfate, sodium chloride, potassium dihydrogen phosphate and sodium malonate (McPherson, 2001) as salts, calcium, zinc, magnesium, manganese and cadmium as divalent cations and the buffers acetate pH 4.5, MES pH 6.0, HEPES pH 7.5 and Tris–HCl pH 8.5. We tested two additives, urea and DMSO, and observed that crystals grown in the presence of urea deteriorated in the crystallization drops over time (after around two weeks). We therefore only retained DMSO as an additive. Similarly, of the three cryoprotectant substances selected initially, glycerol, ethylene glycol and glucose, only one was finally retained, as we did not observe any significant difference between them in crystallization results on test proteins. Following the work of McFerrin & Snell (2002), we decided to use ethylene glycol only.

3.2. Experimental design

We have used the design of experiment and optimization software *MODDE* to prepare the experimental designs. Three successive designs were elaborated and tested, the first two allowing us to refine the number of factors and their useful levels. In the third and final design we used six factors (pH, protein concentration, precipitant, additive, divalent cation and DMSO). These factors were defined as follows.

F_{Hi}. pH is a qualitative factor with four levels (acetate pH 4.5, MES pH 6, HEPES pH 7.5 and Tris–HCl pH 8.5).

F_{Ci}. Protein concentration is a qualitative factor with two levels, low and high (typically, the high concentration was 10 mg ml^{-1}, with 5 mg ml^{-1} for the low concentration).

Table 1
Composition of the 48 solutions in the second kit.

1	0.7 *M* phosphate	2% PEG 400	Blank	0.2 *M* urea	10% ethylene glycol	0.1 *M* HEPES	Low protein
2	1.5 *M* ammonium sulfate	2% MPD	10 m*M* Mg	Blank	10% glucose	0.1 *M* HEPES	Low protein
3	1 *M* malonate	2% 2-propanol	10 m*M* Zn	0.2 *M* urea	10% ethylene glycol	0.1 *M* acetate	Low protein
4	20% PEG 400	0.1 *M* ammonium sulfate	10 m*M* Mg	Blank	10% glucose	0.1 *M* Tris–HCl	Low protein
5	0.7 *M* phosphate	2% MPD	Blank	0.2 *M* urea	10% ethylene glycol	0.1 *M* acetate	Low protein
6	1.5 *M* NaCl	2% 2-propanol	10 m*M* Zn	0.2 *M* urea	10% ethylene glycol	0.1 *M* acetate	Low protein
7	20% PEG 8000	0.07 *M* phosphate	Blank	0.2 *M* urea	10% glucose	0.1 *M* MES	Low protein
8	20% PEG 400	0.1 *M* NaCl	10 m*M* Zn	0.2 *M* urea	10% glycerol	0.1 *M* acetate	Low protein
9	20% 2-propanol	0.1 *M* NaCl	Blank	0.2 *M* urea	10% glycerol	0.1 *M* Tris–HCl	Low protein
10	0.7 *M* phosphate	2% 2-propanol	Blank	5% DMSO	10% ethylene glycol	0.1 *M* HEPES	Low protein
11	1.5 *M* ammonium sulfate	Blank	Blank	Blank	10% glucose	0.1 *M* MES	Low protein
12	20% 2-propanol	0.1 *M* malonate	Blank	0.2 *M* urea	10% glycerol	0.1 *M* Tris–HCl	Low protein
13	20% PEG 8000	0.1 *M* ammonium sulfate	Blank	Blank	10% ethylene glycol	0.1 *M* Tris–HCl	Low protein
14	1 *M* malonate	2% MPD	Blank	5% DMSO	10% glucose	0.1 *M* HEPES	Low protein
15	20% PEG 400	0.07 *M* phosphate	Blank	5% DMSO	10% glycerol	0.1 *M* HEPES	Low protein
16	20% PEG 8000	0.1 *M* malonate	10 m*M* Ca	Blank	10% ethylene glycol	0.1 *M* MES	Low protein
17	1.5 *M* NaCl	2% MPD	10 m*M* Ca	Blank	10% glycerol	0.1 *M* acetate	Low protein
18	20% MPD	0.1 *M* ammonium sulfate	10 m*M* Mg	0.2 *M* urea	10% ethylene glycol	0.1 *M* HEPES	Low protein
19	20% 2-propanol	0.1 *M* ammonium sulfate	10 m*M* Mg	Blank	10% ethylene glycol	0.1 *M* MES	Low protein
20	20% MPD	0.1 *M* NaCl	10 m*M* Zn	Blank	10% glucose	0.1 *M* MES	Low protein
21	1 *M* malonate	Blank	10 m*M* Zn	5% DMSO	10% glucose	0.1 *M* MES	Low protein
22	1 *M* malonate	2% 2-propanol	10 m*M* Mg	Blank	10% glycerol	0.1 *M* Tris–HCl	Low protein
23	20% 2-propanol	0.07 *M* phosphate	Blank	Blank	10% glucose	0.1 *M* MES	Low protein
24	20% 2-propanol	Blank	10 m*M* Ca	5% DMSO	10% glycerol	0.1 *M* HEPES	Low protein
25	20% PEG 400	Blank	10 m*M* Ca	Blank	10% glucose	0.1 *M* acetate	High protein
26	20% PEG 8000	Blank	10 m*M* Mg	0.2 *M* urea	10% glucose	0.1 *M* MES	High protein
27	1.5 *M* NaCl	Blank	10 m*M* Mg	0.2 *M* urea	10% glycerol	0.1 *M* acetate	High protein
28	20% MPD	0.1 *M* malonate	Blank	0.2 *M* urea	10% ethylene glycol	0.1 *M* acetate	High protein
29	1.5 *M* ammonium sulfate	2% PEG 400	Blank	0.2 *M* urea	10% ethylene glycol	0.1 *M* MES	High protein
30	20% MPD	0.07 *M* phosphate	Blank	5% DMSO	10% glucose	0.1 *M* Tris–HCl	High protein
31	20% PEG 400	Blank	10 m*M* Mg	0.2 *M* urea	10% glycerol	0.1 *M* MES	High protein
32	20% MPD	0.1 *M* NaCl	10 m*M* Ca	5% DMSO	10% glycerol	0.1 *M* MES	High protein
33	20% PEG 8000	0.1 *M* NaCl	10 m*M* Mg	5% DMSO	10% glucose	0.1 *M* acetate	High protein
34	0.7 *M* phosphate	2% MPD	Blank	5% DMSO	10% glucose	0.1 *M* HEPES	High protein
35	20% PEG 400	0.1 *M* malonate	10 m*M* Zn	5% DMSO	10% ethylene glycol	0.1 *M* MES	High protein
36	20% MPD	Blank	10 m*M* Zn	5% DMSO	10% ethylene glycol	0.1 *M* MES	High protein
37	1.5 *M* ammonium sulfate	2% PEG 400	Blank	Blank	10% glycerol	0.1 *M* acetate	High protein
38	1.5 *M* ammonium sulfate	2% 2-propanol	10 m*M* Zn	Blank	10% glycerol	0.1 *M* acetate	High protein
39	20% PEG 8000	Blank	Blank	5% DMSO	10% ethylene glycol	0.1 *M* Tris–HCl	High protein
40	20% 2-propanol	0.1 *M* malonate	10 m*M* Mg	5% DMSO	10% glucose	0.1 *M* HEPES	High protein
41	1.5 *M* NaCl	2% PEG 400	10 m*M* Ca	Blank	10% ethylene glycol	0.1 *M* acetate	High protein
42	1 *M* malonate	2% PEG 400	10 m*M* Ca	Blank	10% glycerol	0.1 *M* Tris–HCl	High protein
43	20% PEG 400	0.1 *M* NaCl	10 m*M* Mg	Blank	10% ethylene glycol	0.1 *M* HEPES	High protein
44	20% PEG 8000	0.07 *M* phosphate	Blank	Blank	10% glycerol	0.1 *M* HEPES	High protein
45	20% 2-propanol	0.1 *M* ammonium sulfate	10 m*M* Zn	0.2 *M* urea	10% glucose	0.1 *M* acetate	High protein
46	20% MPD	0.07 *M* phosphate	Blank	Blank	10% glycerol	0.1 *M* Tris–HCl	High protein
47	0.7 *M* phosphate	Blank	Blank	Blank	10% ethylene glycol	0.1 *M* acetate	High protein
48	20% PEG 8000	0.1 *M* NaCl	Blank	Blank	10% glucose	0.1 *M* Tris–HCl	High protein

F_{Pi}. Precipitant is a qualitative factor with eight levels [PEG 8000, PEG 400, MPD, 2-propanol, $(NH_4)_2SO_4$, NaCl, KH_2PO_4 and sodium malonate].

F_{Ai}. Additive is a qualitative factor with eight levels [PEG 400, MPD, 2-propanol, $(NH_4)_2SO_4$, NaCl, KH_2PO_4, sodium malonate and blank].

F_{Ii}. Divalent cation is a qualitative factor with six levels (Ca^{2+}, Cd^{2+}, Mg^{2+}, Mn^{2+}, Zn^{2+} and blank).

F_{Di}. DMSO is a quantitative factor with two levels, 0 and 0.05 [*i.e.* its (*v*/*v*) percentage in the solution].

Ethylene glycol at 15%(*v*/*v*) is a constant component in the experimental design.

All combinations of components known to pose problems of chemical incompatibility (*e.g.* insolubility) were defined as constraints and taken into account. The optimal design proposed by *MODDE* for this set of conditions corresponds to a D-Optimal design (see Tables 1 and 2 for detailed composition of the 48 solutions), allowing only the main effects to be analysed.

3.3. Quantifying the crystallization results

Any statistical evaluation depends on a numerical scale associated with the results of experiments carried out. We therefore devised a scale of 1–10 ranging from precipitates to crystals which was assigned after the end of the experiment. The scale (Q) used is the following. 1, heavy brown or flocculent precipitate; 2, clear drop; 3, phase separation or light precipitate; 4, granular precipitate; 5, microcrystalline precipitate; 6, spherulites; 7, microcrystals; 8, multiple crystals; 9, small crystals (needles, thin plates); 10, well shaped single crystals. We also tested a reduced scale of three levels (regrouping notes of our expanded scale, *i.e.* 1, clear drop or

Table 2
Composition of the 48 screening solutions in the final kit.

1	20% 2-propanol	0.1 *M* ammonium sulfate	10 m*M* Mn	0	0.05 *M* acetate	Low protein
2	20% 2-propanol	0.1 *M* phosphate	Blank	5% DMSO	0.05 *M* acetate	Low protein
3	20% 2-propanol	Blank	10 m*M* Mg	5% DMSO	0.05 *M* HEPES	Low protein
4	1 *M* malonate	Blank	10 m*M* Zn	0	0.05 *M* acetate	Low protein
5	1 *M* malonate	Blank	10 m*M* Ca	5% DMSO	0.05 *M* Tris–HCl	Low protein
6	1 *M* malonate	2% 2-propanol	Blank	5% DMSO	0.05 *M* Tris–HCl	Low protein
7	30% MPD	0.1 *M* malonate	10 m*M* Mn	0	0.05 *M* HEPES	Low protein
8	30% MPD	0.1 *M* phosphate	Blank	5% DMSO	0.05 *M* HEPES	Low protein
9	30% MPD	0.1 *M* malonate	10 m*M* Mg	0	0.05 *M* Tris–HCl	Low protein
10	2 *M* NaCl	2% PEG 400	10 m*M* Ca	5% DMSO	0.05 *M* acetate	Low protein
11	2 *M* NaCl	2% MPD	10 m*M* Zn	5% DMSO	0.05 *M* MES	Low protein
12	2 *M* NaCl	2% 2-propanol	10 m*M* Cd	0	0.05 *M* HEPES	Low protein
13	30% PEG 400	0.1 *M* ammonium sulfate	10 m*M* Mg	0	0.05 *M* acetate	Low protein
14	30% PEG 400	Blank	10 m*M* Cd	0	0.05 *M* MES	Low protein
15	30% PEG 400	0.1 *M* NaCl	Blank	0	0.05 *M* MES	Low protein
16	30% PEG 400	0.1 *M* malonate	10 m*M* Ca	5% DMSO	0.05 *M* MES	Low protein
17	20% PEG 8000	Blank	10 m*M* Mn	5% DMSO	0.05 *M* acetate	Low protein
18	20% PEG 8000	0.1 *M* NaCl	10 m*M* Zn	0	0.05 *M* MES	Low protein
19	20% PEG 8000	0.1 *M* NaCl	10 m*M* Mg	5% DMSO	0.05 *M* Tris–HCl	Low protein
20	0.9 *M* phosphate	2% MPD	Blank	0	0.05 *M* acetate	Low protein
21	0.9 *M* phosphate	Blank	Blank	5% DMSO	0.05 *M* MES	Low protein
22	1.8 *M* ammonium sulfate	2% MPD	10 m*M* Cd	0	0.05 *M* HEPES	Low protein
23	1.8 *M* ammonium sulfate	2% PEG 400	Blank	0	0.05 *M* Tris–HCl	Low protein
24	20% 2-propanol	0.1 *M* NaCl	10 m*M* Cd	5% DMSO	0.05 *M* acetate	Low protein
25	20% 2-propanol	0.1 *M* malonate	Blank	0	0.05 *M* MES	High protein
26	20% 2-propanol	Blank	10 m*M* Ca	0	0.05 *M* Tris–HCl	High protein
27	1 *M* malonate	2% PEG 400	10 m*M* Mg	0	0.05 *M* MES	High protein
28	1 *M* malonate	2% MPD	10 m*M* Mn	5% DMSO	0.05 *M* MES	High protein
29	1 *M* malonate	2% MPD	10 m*M* Ca	0	0.05 *M* HEPES	High protein
30	1 *M* malonate	2% PEG 400	10 m*M* Cd	5% DMSO	0.05 *M* HEPES	High protein
31	30% MPD	0.1 *M* NaCl	1 0m*M* Ca	0	0.05 *M* acetate	High protein
32	30% MPD	Blank	Blank	5% DMSO	0.05 *M* acetate	High protein
33	30% MPD	Blank	10 m*M* Zn	0	0.05 *M* MES	High protein
34	30% MPD	0.1 *M* ammonium sulfate	10 m*M* Cd	5% DMSO	0.05 *M* MES	High protein
35	2 *M* NaCl	Blank	Blank	0	0.05 *M* acetate	High protein
36	2 *M* NaCl	Blank	10 m*M* Mn	0	0.05 *M* MES	High protein
37	2 *M* NaCl	2% MPD	10 m*M* Mg	5% DMSO	0.05 *M* Tris–HCl	High protein
38	30% PEG 400	0.1 *M* malonate	10 m*M* Zn	5% DMSO	0.05 *M* acetate	High protein
39	30% PEG 400	0.1 *M* NaCl	10 m*M* Mn	5% DMSO	0.05 *M* HEPES	High protein
40	30% PEG 400	0.1 *M* phosphate	Blank	0	0.05 *M* Tris–HCl	High protein
41	20% PEG 8000	0.1 *M* malonate	10 m*M* Cd	0	0.05 *M* acetate	High protein
42	20% PEG 8000	0.1 *M* phosphate	Blank	0	0.05 *M* MES	High protein
43	20% PEG 8000	0.1 *M* ammonium sulfate	Blank	5% DMSO	0.05 *M* HEPES	High protein
44	0.9 *M* phosphate	2% 2-propanol	Blank	0	0.05 *M* HEPES	High protein
45	0.9 *M* phosphate	2% PEG 400	Blank	5% DMSO	0.05 *M* Tris–HCl	High protein
46	1.8 *M* ammonium sulfate	2% 2-propanol	10 m*M* Mg	5% DMSO	0.05 *M* acetate	High protein
47	1.8 *M* ammonium sulfate	2% 2-propanol	10 m*M* Mn	5% DMSO	0.05 *M* MES	High protein
48	1.8 *M* ammonium sulfate	Blank	10 m*M* Mg	5% DMSO	0.05 *M* HEPES	High protein

precipitate; 2, crystalline precipitate; 3, crystals) as suggested by Abergel *et al.* (1991).

3.4. Evaluation of the results

The D-Optimal design chosen uses the following model, where Q is a function defined by a linear combination of the factors F_{ji},

$$Q_{ij} = \text{constant} + \sum \alpha_i F_{ji},$$

and all trials (j) within the design form a matrix. The *MODDE* software allows the use of multiple least-squares regression (MLR) or partial least-squares (PLS) methods to derive the best values of the coefficients α_i. Both methods have been shown to be applicable to crystallization experiments (Carter, 1990; Sedzik, 1995; Sedzik & Norinder, 1997). Unlike PLS, the MLR method is based on the assumption that all factors are independent.

3.5. Optimization of the results

The crystallization screening process will rarely give crystals that are directly suitable for diffraction studies, although there may be exceptions. When using standard commercial kits, it is usually necessary to modify by trial and error the concentration of each of the compounds used in the kit solution that gave an interesting result. If no result is obtained, then the crystallographer has to turn to other kits, repeating the screening process. In the experimental design approach, on the other hand, the results should give leads concerning the compounds that are favourable for a given protein and that can be optimized.

We propose two approaches for the optimization of results obtained during the first screening stage. If the indications given by the statistical treatment of the results point clearly towards a small number of conditions (*e.g.* pH range, single precipitant, divalent ion), as well as components to avoid, then the *MODDE* (or any other) software can be used for a full factorial design which will sample the selected conditions and their levels. In this case it should be possible to obtain a good statistical model of the system, which will allow proper response surface modelling, thus locating the best set of conditions to use. The test case of trypsin illustrates this approach (Fig. 1).

Most frequently, however, the results will not be clear enough to limit the optimization design to a sufficiently small number of factors for a full factorial design. Indeed, the statistical model validity is not always very good, as illustrated in Fig. 2(*a*). The coefficient plot, however (Fig. 2*d*), does give a clear and significant indication of conditions that are favourable for the protein under study. The strategy then is to use a fractional factorial design, sampling the conditions indicated by the screening result. The result of the fractional factorial design should allow the evaluation of the factors and some of their interactions. A final optimization run could then be designed to establish the target-values domain of the influential factors.

We also propose an alternative approach to optimization, intended in particular for proteins that do not yield clear results with the initial screen. The factors indicated as positive

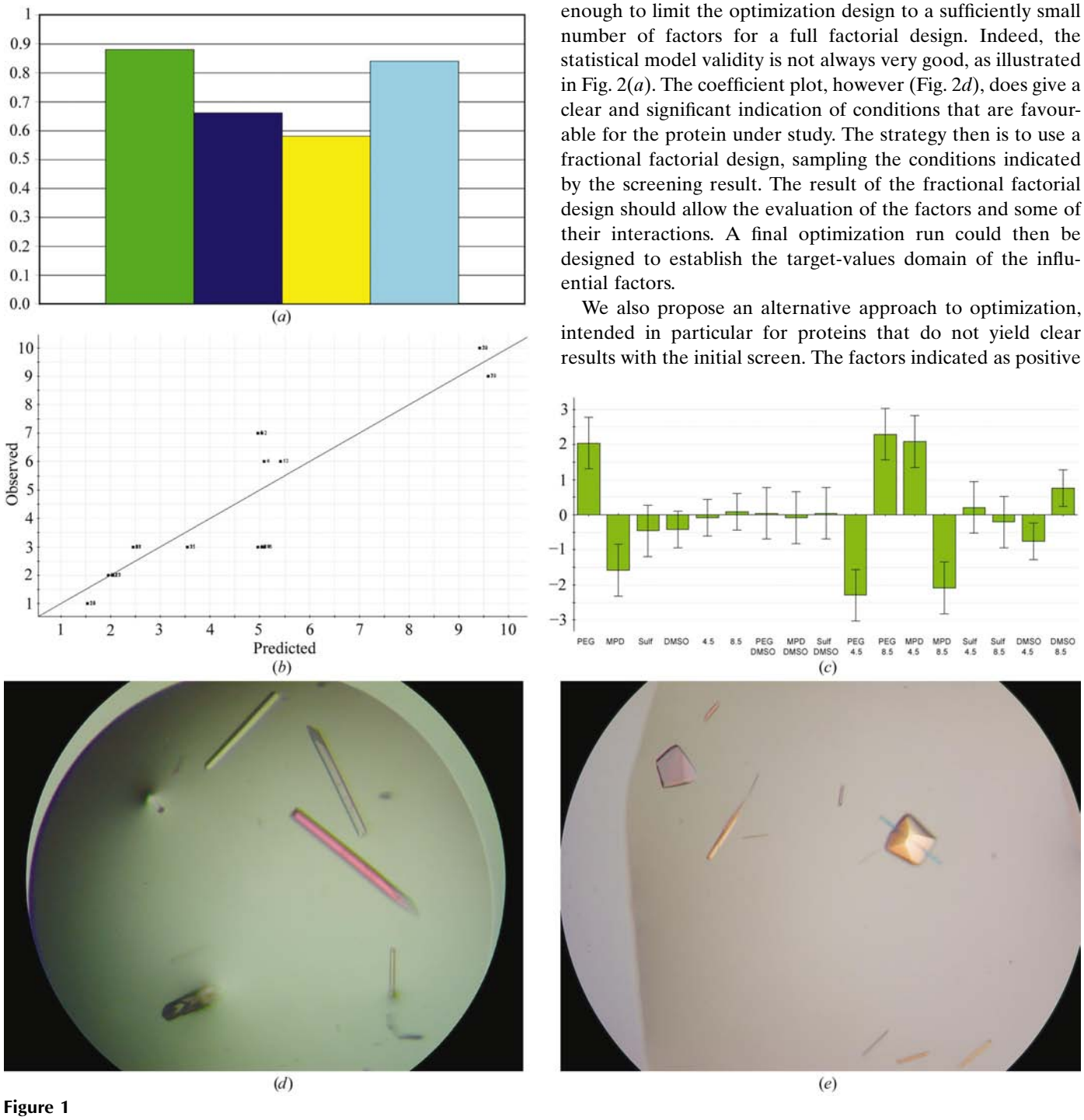

Figure 1
Example of the first optimization strategy, the case of trypsin. The screening of crystallization conditions for trypsin gave the following leads: F_{P1} = PEG 8000, F_{P2} = MPD and F_{P3} = ammonium sulfate as possible precipitants at pH either F_{H1} = 4.5 or F_{H2} = 8.5, with F_D = DMSO as a positive factor. The optimization was carried out using a full factorial design combining those factors. (*a*) Summary plot (R^2, green, is the fraction of variation of the response explained by the linear model; Q^2, blue, is the fraction of variation of the response that can be predicted by the model; *V*, yellow, is a measure of the model validity; *P*, cyan, is a measure of reproducibility of the results); (*b*) plot of observed against predicted results, validating the statistical model; (*c*) coefficient (α_i) plot; (*d*) crystals obtained with PEG 8000 at pH 8.5 and (*e*) with 5% DMSO added, respectively.

in the initial screen are retained and we test, in addition, a detergent (0.05% β-octyl glucoside), an organic compound (5% 1,6-hexanediol) and an amino acid (0.1 *M* L-arginine). We applied this approach to the optimization of phospholipase A_2 crystallization conditions. The initial screen gave the following leads: NaCl, PEG 8000 and 2-propanol as precipitants at pH 4.5 and 6.0, with phosphate, calcium and magnesium ions as additives. The results were not sufficiently clear-cut and an optimization of 24 experiments using a D-Optimal design and including the three additives mentioned above was carried out. All results showed calcium to be important, while β-octyl glucoside was a positive factor, allowing a full factorial design to refine the crystallization conditions.

4. Discussion

Our aim was to apply the principles of statistical experimental design to the screening of protein crystallization conditions, following on from the work of Carter (1990), Abergel *et al.* (1991) and others, to the use of preformulated general sets of solutions (kits). The kit that we arrived at corresponds to a design that is satisfactory from the statistical point of view, given the number of factors sampled, the number of experiments chosen and the constraints imposed by chemical compatibilities of the solution components used. The only design possible is a D-Optimal design, which was used in all kits tested. The final design used has good efficiency (72%, compared with 100% for a full fractional design) and determinant value of the information matrix (LogD = 25) [for definitions see Box *et al.* (1978) or the NIST/SEMATECH *e-Handbook of Statistical Methods* at http://www.itl.nist.gov/div898/handbook/].

The analysis of the result is reliable and its quality can be evaluated by standard statistical tools, such as goodness of fit, analysis of variance, analysis of residuals *etc.* The analysis of the crystallization results is, however, very sensitive to the quality scale used. We have found that in the more general case where a protein does not give reasonable crystals, different types of precipitate have to be graded very carefully, while different types of crystals need not be given many different grades. Thus, rather than follow the quality scale used by Carter & Carter (1979), which allows only three grades for precipitates and semi-crystalline objects compared with six grades for different crystal morphologies, we decided to expand the range of grades for different types of precipitate and to include clear drops, while reducing the range for different crystal types.

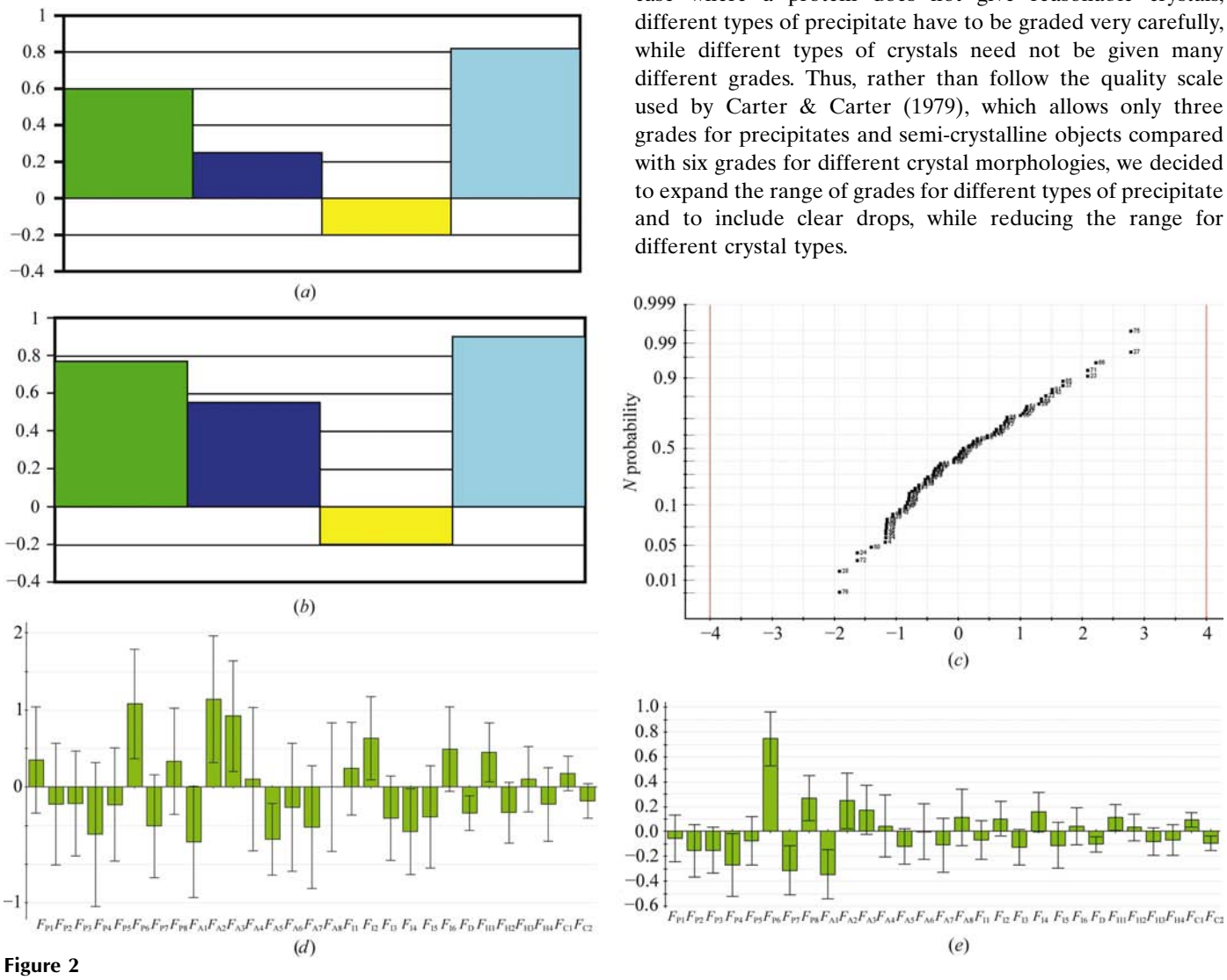

Figure 2
Example of statistical evaluation of the screening results for yeast ORF 116 (YER067w): (*a*) summary plot before and (*b*) after the removal of trials whose residuals are large (the colour code is as in Fig. 1); (*c*) a plot of residuals corresponding to (*a*), (*d*) the coefficient (α_i) plot of the same experiment, using the full and (*e*) the reduced quality scale. The effect of Mg^{2+} (F_{I2}) would be missed using the reduced scale.

Table 3
Results of a comparison of our kit against Hampton Crystal Screens I and II: number of conditions that gave crystals.

The kit used for this comparison was the second one we elaborated.

Protein	Our kit	Crystal Screens I and II
α-Chymotrypsin	5	8
Human apotransferrin	2	1
BSA	2	0
Glucose isomerase	8	17
Insulin	23	12
Lysozyme	11	9
Papain	5	2
Triosephosphate isomerase	0	1
Trypsin	1	1
Ubiquitin	0†	2
Xylanase	2	2
Grd19p	2	2
YIL020c	2	1
YER067w	2	3
YD2435c	0	1

† The successful Hampton Screen conditions contained Cd^{2+} ions.

The results fall into two categories: either there were crystalline forms obtained in at least some conditions or the trials gave only precipitates of different types. In most cases, the initial analysis gives at least qualitative answers. The simplified quality scale of three points mentioned above: (1, clear drop or precipitate; 2, crystalline precipitate; 3, crystals) cannot be used in cases where the screening did not give rise to some kind of crystals. It is, however, useful in more clear-cut cases, where it makes the analysis easier without deteriorating the results too much (Figs. 2*d* and 2*e*). This is of particular interest for the development of automatic analysis of crystallization screening results, especially as most image-analysis algorithms so far have had difficulties in defining subtle differences in the aspect of crystallization drops (Wilson, 2002).

In order to evaluate the overall efficiency of our set of crystallization conditions, we made a comparison with Hampton Crystal Screens I and II for a number of proteins. The success rate for the proteins tested is similar (Table 3), thus validating our choice. On the basis of these results, one could suggest that the propensity for crystallization of a protein is defined by fundamental solution properties such as pH, the presence of an appropriate precipitating agent and certain divalent ions (which often occur as natural ligands anyway). The final crystal quality, on the other hand, may be influenced by other solution components, which in our approach would be searched for in the second, optimization stage.

We do not see any deterioration of crystallization success caused by the inclusion of a cryoprotecting agent in the solution formulations. The test proteins gave crystals with a similar success rate as when standard commercial screens were used. While we have not been able to test the actual cryoprotecting efficiency of all the solutions, it is likely they should be adequate using flash-cooling (McFerrin & Snell, 2002). Since a cryoprotectant compound is used systematically, it could be a good idea to add it directly to the protein solution, thus permitting its safe storage by freezing as an added advantage. This would also allow increasing the concentrations of some preciptating agents whose solubility is a limiting factor in the solution formulations.

5. Conclusions

We have tried to combine the versatility of the use of crystallization kits that explore a wide range of conditions with the rigorous methodology of statistical experimental designs. We show that while the efficiency of locating the conditions in which a given protein should crystallize is not compromised by the smaller number of compounds tested, the method provides a much more straightforward and efficient way for the subsequent optimization of these conditions. We propose two different methods of optimization, depending on the screening results. We think that our approach is suitable for use in the context of automation of protein crystal growth, while allowing further improvements in the future. Finally, we show that it is possible to systematically include a cryoprotecting agent within the crystallization solution formulation.

This work was carried out as part of a 'Genhomme' project financed by the French Ministère de Recherche et Technologie for the development of an automated crystallization platform. TTT and IS were financed by the MRT grant.

References

Abergel, C., Moulard, M., Moreau, H., Loret, E., Cambillau, C. & Fontecilla-Camps, J. C. (1991). *J. Biol. Chem.* **266**, 20131–20138.
Audic, S., Lopez, F., Claverie, J. M., Poirot, O. & Abergel, C. (1997). *Proteins Struct. Funct. Genet.* **29**, 252–257.
Box, E. P., Hunter, W. G. & Hunter, J. S. (1978). *Statistics for Experimenters.* New York: John Wiley & Sons.
Bukrinsky, J. T. & Navarro-Poulsen, J. C. (2001). *J. Appl. Cryst.* **34**, 533–534.
Carter, C. W. (1990). *Methods*, **1**, 12–24.
Carter, C. W. Jr & Carter, C. W. (1979). *J. Biol. Chem.* **254**, 12219–12223.
Chayen, N. E. & Saridakis, E. (2002). *Acta Cryst.* D**58**, 921–927.
Garman, E. (1999). *Acta Cryst.* D**55**, 1641–1653.
Jancarik, J. & Kim, S.-H. (1991). *J. Appl. Cryst.* **24**, 409–411.
McFerrin, M. B. & Snell, E. H. (2002). *J. Appl. Cryst.* **35**, 538–545.
McPherson, A. (2001). *Protein Sci.* **10**, 418–422.
Majeed, S., Ofek, G., Belachew A., Huang, C. C., Zhou, T. & Kwong, P. D. (2003). *Structure*, **11**, 1061–1070.
Page, R., Grzechnik, S. K., Canaves, J. M., Spraggon, G., Kreusch, A., Kuhn, P., Stevens, R. C. & Lesley, S. A. (2003). *Acta Cryst.* D**59**, 1028–1037.
Sedzik, J. (1995). *Biochim. Biophys. Acta*, **1251**, 177–185.
Sedzik, J. & Norinder, U. (1997). *J. Appl. Cryst.* **30**, 502–506.
Segelke, B. W. (2001). *J. Cryst. Growth*, **232**, 553–562.
Wilson, J. (2002). *Acta Cryst.* D**58**, 1907–1914.
Wooh, J. W., Kidd, R. D., Martin, J. L. & Kobe, B. (2003). *Acta Cryst.* D**59**, 769–772.
Zhou, C. Z., Li de la Sierra-Gallay, I., Quevillon-Cheruel, S., Collinet, B., Minard, P., Blondeau, K., Nenckes, G., Aufrère, R., Leulliot, N., Grailles, M., Sorel, I., Savarin, P., de la Torre, F., Poupon, A., Janin, J. & van Tibeurgh, H. (2003). *J. Biol. Chem.* **278**, 50371–50376.