Experimental optimization of protein refolding with a genetic algorithm

Bernd Anselment,¹ Danae Baerend,² Elisabeth Mey,² Johannes Buchner,² Dirk Weuster-Botz,¹ and Martin Haslbeck^{2*}

¹Lehrstuhl für Bioverfahrenstechnik, Technische Universität München, Boltzmannstr. 15, D-85748 Garching, Germany

²Department Chemie and Center for Integrated Protein Science Munich (CIPSM), Technische Universität München, D-85748 Garching, Germany

Received 20 May 2010; Revised 16 August 2010; Accepted 17 August 2010

DOI: 10.1002/pro.488

Published online 26 August 2010 proteinscience.org

Abstract: Refolding of proteins from solubilized inclusion bodies still represents a major challenge for many recombinantly expressed proteins and often constitutes a major bottleneck. As *in vitro* refolding is a complex reaction with a variety of critical parameters, suitable refolding conditions are typically derived empirically in extensive screening experiments. Here, we introduce a new strategy that combines screening and optimization of refolding yields with a genetic algorithm (GA). The experimental setup was designed to achieve a robust and universal method that should allow optimizing the folding of a variety of proteins with the same routine procedure guided by the GA. In the screen, we incorporated a large number of common refolding additives and conditions. Using this design, the refolding of four structurally and functionally different model proteins was optimized experimentally, achieving 74–100% refolding yield for all of them. Interestingly, our results show that this new strategy provides optimum conditions not only for refolding but also for the activity of the native enzyme. It is designed to be generally applicable and seems to be eligible for all enzymes.

Keywords: protein refolding; inclusion bodies; high-throughput optimization; genetic algorithms; multiobjective optimization

Introduction

Despite extensive experimental and theoretical work, refolding of proteins from inclusion bodies

remains a major bottleneck in supplying recombinant proteins for research and industrial applications.^{1–4} A basic problem in this context is the complexity of the refolding process and the broad variety of influencing parameters. Importantly, a specific experimental parameter screening is necessary for each protein. Several reviews address this issue^{1–5} and experimental guidelines⁶ for initial refolding experiments are available. In this context, rational high throughput refolding strategies have been coming into focus. In the last decade, several screening procedures were developed that search for conditions that allow refolding of proteins.^{7–14} However, these approaches are limited in some core characteristics. They either analyze only a limited collection of buffer components that affect refolding or fail to sufficiently consider their interdependence. Another shortcoming of previous studies is the absence of further optimization of suitable refolding conditions. Therefore, a combined approach of

Abbreviations: DTT, dithiothreitol; EDTA, ethylenediaminetetraacetic acid; GA, genetic algorithm; Gdn-HCl, guanidine hydrochloride; gen, generation; GFP, green fluorescent protein; GLK, glucokinase; GLR, glutathione reductase; GSH, L-glutathione reduced; GSSG, L-glutathione oxidized; HEPES, 4-(2-hydroxyethyl)piperazine-1-ethanesulfonic acid; LYZ, lysozyme; MOPS, 4-morpholinopropanesulfonic acid; PEG 4000, polyethylene glycol 4000; phosphate buffer, sodium phosphate buffer; SPEA 2, strength pareto algorithm; TCEP, tris (2-carboxyethyl) phosphine hydrochloride; Tris-HCl, 2-amino-1,3-dihydroxy-2-(hydroxymethyl)-propane.

Grant sponsor: International Graduate School of Science and Engineering (IGSSE), Technische Universität München.

*Correspondence to: Martin Haslbeck, Lehrstuhl für Biotechnologie, Department Chemie, Technische Universität München, D-85748 Garching, Germany.
E-mail: martin.haslbeck@ch.tum.de

screening and optimization is desirable. To overcome these basic shortcomings, we developed screening and optimization procedures for protein refolding by dilution from the completely unfolded state utilizing a genetic algorithm (GA).

Evolutionary stochastic search and optimization strategies like GA have gained popularity since Rechenberg and Holland^{15,16} published their fundamental work in the 1980s. GAs simulate the process of natural evolution starting with a set of randomly generated candidates, which iteratively evolve to better solutions during the optimization. Two basic evolutionary principles are used: selection and variation. Although selection imitates the competition for reproduction among organisms, variation is the capability to generate new solutions via recombination and mutation.

This stochastic search algorithm of generations can be summarized as follows for an experimental application like protein refolding. In a first step, the problem is defined (objectives) and critical parameters are identified. This information is used to define the search space, that is, what experimental conditions should be tested. In the second stage of GA implementation, an iterative series of experiments is performed. For each experiment, the GA proposes a collection of unique experimental conditions (generation) that is subsequently evaluated experimentally. While the first experiment contains random experimental conditions, afterward the principles of evolution are applied. The GA selects the “best” conditions in each experiment with highest probability and varies these with mutation operators to obtain related new conditions. Subsequently, these new conditions are evaluated in the next parallel experiment. This process is continued (until terminated by the user), resulting in a heuristic optimization strategy that mimics the process of natural evolution.

GAs are considered as robust and powerful search and optimization methods especially for large complex search spaces and multiple objectives.¹⁷ Therefore, they should be very well suited for a complex problem like protein refolding, where numerous parameters and substances affect the outcome. GAs are widely used in informatics; furthermore, they have been successfully applied for a wide variety of experimental problems. Recent examples include processes in chemistry^{18,19} and reaction media optimization in biochemical engineering.^{20,21} An important point is the stochastic nature of the optimization. Because of limited number of experiments possible during optimization, a precise identification of the global optimum (e.g., the best refolding conditions) is improbable. However, GAs offer the chance to efficiently analyze large and complex search spaces with limited experimental effort. Significant improvements of the optimization objectives can be acquired with relatively low experimental effort.²¹

Implementation of a GA for multiobjective experimental optimization was recently demonstrated²² and, thus, offers the chance for further reduction of the experimental effort.

Here, we adapted this concept to the experimental optimization of protein refolding yields. Especially, as the possibility to optimize for two objectives in parallel seemed highly interesting, we decided to employ a GA allowing a multiobjective optimization. This offered the chance to optimize not only the refolding yields but also the underlying activities or costs of the refolding reaction in parallel. To our best knowledge, this is the first time that a GA was used to approach the screening for protein refolding and its yield optimization in a single process. Our work describes a robust and simple experimental setup that proved to be applicable for four structurally different model proteins.

Results

Experimental design

To establish a new approach to optimizing protein folding conditions, we adjusted a GA and defined a suitable experimental setup (Fig. 1). The optimization process is iterative and one iteration (one set of experiments) is referred to as a generation (gen). For each protein, the optimization started with a random collection of possible solutions, i.e., 22 variations of refolding conditions (first gen). Each variation was evaluated by diluting the denatured protein into the respective refolding condition. The success in terms of the optimization objectives (e.g., refolding yields, costs, or activities) was evaluated. In our experiments, we optimized two objectives at the same time: (a) refolding yields and costs of the refolding condition or alternatively (b) the activities of the native and refolded protein in the respective refolding condition. Similar to evolution, experimental conditions with high fitness (i.e., efficient refolding, low costs, and high activities) were then selected to finally calculate a new set of refolding conditions (second gen). These new refolding conditions, basing on the most efficient solutions of the previous gen, again were evaluated experimentally. Within several gens, a gradual shift toward more efficient refolding occurred. The optimization was terminated, if no increase in performance was determined in several generations or a fixed limit of experiments was reached.

As a GA has never been used before to investigate protein refolding, a configuration of the algorithm to the specific experimental problem was required. The decisive factor for the optimization success is the design of the refolding conditions that are substances which should be tested in the refolding experiments at what concentrations and in which combinations? In the last decades, a large

Figure 1. Scheme of the optimization strategy using a GA. Twenty-two refolding conditions were randomly generated (first gen) and each of them was subsequently evaluated experimentally. Depending on the results, fitness values were assigned to each condition. The best conditions were selected and 22 new conditions were generated by the algorithm; these were again tested experimentally. This iterative optimization can be continued until no further improvement of the fitness values is detected in several gens.

number of conditions, substances, and methods for protein refolding have been established.^{1–14} Experimental variables and parameters were extracted from the refolding literature and combined with the information on ~1000 refolding experiments from the REFOLD database^{23,24} to establish a comprehensive experimental design (summarized in Table I).

Functionally related substances and conditions were subgrouped in six different classes. (a) The first class refers to the pH value and the buffer substance. Interestingly, 2-amino-1,3-dihydroxy-2-(hydroxymethyl)-propane (Tris-HCl) and sodium phosphate buffers (phosphate buffers) were most prominent in REFOLD.^{23,24} Nevertheless, we also included 4-(2-hydroxyethyl)piperazine-1-ethanesulfonic acid (HEPES) and 4-morpholinepropanesulfonic acid (MOPS), two other common organic buffers. Concentrations varied between 20 and 100 mM for phosphate, HEPES, and MOPS and up to 1.25M for Tris-HCl, because it is also employed as a refolding additive.¹ pH values between 6.0 and 9.5 cover most refolding experiments.^{23,24} Only conditions inside the buffer range were allowed as follows: phosphate buffer, pH 6.0–7.5 and Tris-HCl, pH 7.0–9.5. (b) NaCl was used as the primary compound to vary the ionic strength of the buffer. Furthermore, addition of small concentrations (20 mM) of KCl

was tested. (c) Another parameter class was composed of refolding additives, including glycerol and polyethylene glycol 4000 (PEG 4000) as well as three commonly used amino acids. (d) Divalent metal cations that are used in past refolding experiments¹⁴ and alternatively ethylenediaminetetraacetic acid (EDTA) have formed the fourth parameter class. (e) Also detergents, including different detergent families (zwitterionic, ionic, and nonionic) in concentrations between 0 and 4/3 of the critical micelle concentration, were incorporated as one parameter class. In addition, a nondetergent sulfobetaine that was previously used in refolding screens^{9,25} was included. (f) As disulfide bonds play a central role in protein structure determination, common reducing and oxidizing agents [dithiothreitol (DTT), tris (2-carboxyethyl) phosphine hydrochloride (TCEP), L-glutathione reduced (GSH), and L-glutathione oxidized (GSSG)] formed the last class.

On this basis, an important issue is the coding of the six parameter classes. The minimal buffer is defined only by its pH and the corresponding buffer substance (e.g., Tris-HCl, pH 8.0). To allow a highly flexible variation of all parameters, addition of components from all other parameter classes to this minimal buffer was possible, resulting in complex refolding conditions like Tris-HCl, pH 8.0 with 100

Table I. *Experimental Design of the Proposed Refolding Optimization Strategy*

Parameter/substance	Min	Max	Unit
pH	6.0	9.5	—
Buffer substances		No combination	
Phosphate buffer	20	100	mM
HEPES	20	100	mM
MOPS	50	100	mM
Tris-HCl	20	1250	mM
Salts		Combination: NaCl and KCl	
NaCl	0	350	mM
KCl	0	20	mM
Additives	Combination: (glycerol or PEG) and (arginine and glutamine and glycine)		
Glycerol	0	15	% v/v
PEG	0	0.2	% w/v
Arginine	0	750	mM
Glutamine	0	100	mM
Glycine	0	150	mM
Cofactors		No combination	
(Cu ²⁺ , Zn ²⁺ , Mg ²⁺ , and Mn ²⁺)	0	5	mM
EDTA	0	2	mM
Detergents		No combination	
Zwitterionic			
CHAPS	0	10.67	mM
Zwittergent 3–12	0	4	mM
NDSB 201	0	1500	mM
Nonionic			
Tween 20	0	0.08	mM
Triton X-100	0	0.8	mM
Brij 35	0	0.12	mM
Ionic			
SDS	0	12	mM
Sodium desoxycholate	0	8	mM
Redox agents		Combination: DTT or TCEP or (GSH and GSSG)	
DTT	0	10	mM
TCEP	0	10	mM
GSH	0	5	mM
GSSG	0	5	mM

(Cu²⁺, Zn²⁺, Mg²⁺, Mn²⁺), sulfates; CHAPS, 3-((3-Cholamidopropyl)dimethylammonium)-1-propanesulfonate; Zwittergent 3–12, 3-(dodecyldimethylammonio)-propanesulfonate; NDSB 201, nondetergent sulfobetaine 201, no detergent but grouped in this functional class; Tween 20, polyethylene glycol sorbitan monolaurate; Triton-X 100, polyethylene glycol tert-octyl-phenyl ether; Brij 35, polyethylene glycol dodecyl ether; SDS, sodium dodecyl sulfate; sodium desoxycholate, 7-deoxycholic acid sodium salt.

The table lists minimal and maximal values for the respective parameters and substances, sorted in functional classes. A refolding condition consisted of at least one pH and one buffer substance (e.g., Tris-HCl, pH 8.0). Additionally, substances from other classes could be included (e.g., Tris-HCl, pH 8.0 with 100 mM NaCl, 100 mM arginine) or left out. Furthermore, combinations within several classes were possible. These were annotated with “and” and not possible combinations with “or.”

mM NaCl, 100 mM arginine, and 5 mM DTT. Furthermore, to screen for synergistic interactions, combinations inside one functional class were allowed in several cases, for example, both glutamine and arginine as refolding additives.²⁶ Table I summarizes detailed information on the experimental design, forbidden combinations are annotated with “or” and possible combinations with “and.” To design a robust and technically easy to handle optimization strategy, the refolding method, the refolding temperature, and the final protein concentration were standardized to 10°C and 5 µg mL^{−1}, respectively.

Refolding optimization of model proteins

To assess the performance of the proposed optimization strategy, well characterized model enzymes

were chosen for the experiments. All proteins were available both in the denatured and soluble, native form. We used enzyme-specific functional assays to exactly quantify the refolding success. To exclude effects of the various refolding additives on the functional assays, the refolding yield was quantified individually for each refolding condition. Both denatured and native proteins were diluted into the respective refolding conditions. Afterward, refolding yields were calculated as the ratio of the activity of the refolded protein and the native protein in the respective refolding condition. Table II gives an overview of the analyzed proteins and their characteristics.

First, we used the GA to optimize both the refolding yields of green fluorescent protein (GFP)

Table II. Overview of Analyzed Proteins

Abbreviation	Protein	M [kDa]	pI	qs	ds	Activity	Organism
GFP	Enhanced green fluorescent protein	28	5.7	Monomer	–	Intrinsic fluorescence	<i>Aequorea victoria</i>
GLK	Glucokinase	35	6.1	Dimer	–	Phosphorylation of glucose	<i>Escherichia coli</i>
GLR	Glutathione reductase	53	7.7	Dimer	–	Reduction of glutathione disulfide	<i>Saccharomyces cerevisiae</i>
LYZ	Lysozyme	14	9.3	Monomer	4	Hydrolysis of peptidoglycan linkages	<i>Gallus gallus</i>

qs, quaternary structure; ds, disulfide bonds.

and glutathione reductase (GLR) taking into account not only the yield but also the cost of the refolding condition. Figure 2 shows an overview of the optimization approaches for GFP and GLR. The general aim of the optimization was to find refolding conditions with high yields and low costs, that is, conditions that lie in the upper right corner of the graphs. Both optimizations performed remarkably well, obtaining 100% refolding yield for each protein. In case of GFP, the maximum yield was reached after 4 gens; for GLR, already the first gen contained a buffer with 100% yield. However, this buffer was expensive (0.075 € mL⁻¹). Naturally, the optimizations could have been terminated at this point, if yield would have been the only selection parameter. As we also wanted to minimize the respective costs of the refolding condition, we continued the experiments up to gen 6 leading to improved refolding conditions with 100% yield and reduced costs (0.025 € mL⁻¹).

Especially, striking was the fast optimization of the yield for GFP and GLR and for the fact that we obtained good yields with a variety of refolding conditions. Consequently, many of the chosen parameters might have no or little effect on the refolding success, or we were unable to identify positive effects because we already achieved maximum yield. One reason for the second interpretation is the relative yield calculation. For example, GLR resulted in many refolding conditions with 100% yield; however, the respective activities itself vary between 40 and 100 U mg⁻¹, depending on the respective buffer condition (Fig. 3).

In this respect, the (enzymatic) activity of the refolded protein under the refolding conditions seemed to be a good quality attribute to further differentiate our results. Of course, as higher activities are desirable, we set out to optimize both parameters in parallel to obtain buffers with highly active protein and maximum refolding yield (100% of the

Figure 2. Results of the optimization approaches with GFP (A) and GLR (B). Refolding yields and the costs of the refolding buffer were optimized in parallel. Experimental data of the individual gen (1 ●, 2 ○, 3 ▼, 4 △, 5 ■, 6 □) is plotted according to the two objectives, only conditions with costs smaller than 0.05 € mL⁻¹ are displayed. The stars represent an experimentally verified standard refolding condition for each protein.^{26,27} The optimization progress from the start (gen 1) to the end (gen 6) is highlighted for several gens by the black dashed lines. The lines correspond to the Pareto fronts, the best refolding conditions obtained up to and including the current Gen. Consequently, if no better conditions exist the Pareto front remains unchanged and the optimization is not progressing. For the final gen (in brackets: last gen with changes), the experimental error of the Pareto front is highlighted with grey shading.

Figure 3. Specific activities of the native (black) and refolded (grey) GLR in three different refolding conditions (A, B, and C). All conditions showed ~100 % refolding yield but different specific activities. (A) 1M Tris-HCl, pH 8.5, 150 mM NaCl, 10% v/v glycerol, 500 mM arginine, 100 mM glutamine, 2 mM EDTA; (B) 100 mM phosphate buffer, pH 7.5, 250 mM NaCl, 20 mM KCl, 500 mM arginine, 100 mM glutamine, 2 mM EDTA, 5 mM GSH; (C) 100 mM MOPS, pH 8.5, 150 mM NaCl, 20 mM KCl, 500 mM arginine, 50 mM glutamine, 5 mM EDTA.

native activity). We tested this new strategy (optimize native and refolded activity instead of yield) in a second, independent approach for GLR and with two additional proteins: glucokinase (GLK) and lysozyme (LYZ). To avoid limitations, the costs were not optimized in these experiments (Fig. 4).

This second optimization strategy (the parallel optimization of native and refolded activities) was successful for all proteins (Fig. 4). The goal to achieve high native and refolded activities (upper right corner of Fig. 4) and good refolding yields (data points close to the dashed line in Fig. 4) was reached for in all three cases.

For GLR, we again easily reached 100% refolding yield [Fig. 4(A)]. The specific activities obtained were comparable to those obtained using the first optimization approach. This indicates that the highest activities (100–120 U mg⁻¹) were already detected in the first approach. Overall, this second, independent optimization approach resulted in similar optimum refolding conditions compared with the first approach (Fig. 3, data not shown). For GLK, we also achieved maximum yields and, strikingly, a steady increase of the activities during the optimization (from ~65 U mg⁻¹ in gen 1 to ~266 U mg⁻¹ in gen 8). Thus, both objectives were successfully optimized in parallel [Fig. 4(B)].

In a further independent optimization approach, the refolding of the model protein LYZ was analyzed [Fig. 4(C)]. This is a challenging model protein that needs to be oxidized correctly to regain its activity. To assess whether the algorithm is able to select conditions for the correct formation of disulfide

bridges, LYZ was completely reduced and denatured (addition of 5 mM DTT during denaturation). Even after 4 gens (88 refolding conditions), positive results were very sparse indicating that the optimization was not progressing [Fig. 4(C)]. However, the analysis of the experimental data showed a clear trend toward oxidative refolding conditions. The two refolding conditions that showed significant refolding contained both 0.5 mM GSSG and no reducing agent. Thus, even if we had not known that we needed oxidizing conditions, the trend recognized by the algorithm would have led the educated user to this interpretation. As the combinatorial setup of the algorithm prefers reducing conditions, which is quite effective in general but seems to limit the refolding of disulfide-bridged proteins, we modified the configuration of the GA and allowed only GSH/GSSG as redox agents (excluding reductive conditions with DTT and TCEP). In this second optimization approach, despite a large fraction (~80%) of completely inefficient conditions with no refolding yields and no native activity, the optimization nevertheless achieved efficient refolding conditions [Fig. 4(D)]. Thus, overall, the optimization with the GA allowed recognition of the bias toward oxidizing conditions and was robust enough to select highly specific conditions within a predominantly negative background.

To assess the performance of the described optimization approaches, a comparison with the literature is necessary. Table III displays the performance and composition of the two optimum refolding conditions and an experimentally verified literature reference.^{26–28} The experimental verification of the literature reference was necessary because factors like refolding method, time, temperature, and protein concentration might significantly affect the refolding success.^{1–5}

Before interpreting the results of Table III, it is important to note that the four protein references (R) were not used (as a starting point) for the optimization. They were only a comparison with literature refolding conditions using the same refolding method.

For GFP, the optimization approach resulted in two refolding conditions with ~100% yield, whereas Ref. 26 achieved only 87%. However, it should be noted that the determination of GFP fluorescence and thus the yield showed high relative errors of up to 50% depending on the buffer components. Nevertheless, the two optimum refolding conditions derived from our optimization approach were quite similar to the reference condition.²⁶ They include a neutral or slightly alkaline pH of 7.0 or 7.5, as well as NaCl, a combination of arginine and glutamine and a reducing component. The difference lies mainly in the relative concentrations and the presence of glycerol.

Figure 4. Results of the optimization approaches with GLR (A) and GLK (B) and LYZ (C and D). Native and refolded activities were optimized in parallel. Experimental data of the individual gen (1 ●, 2 ○, 3 ▼, 4 △, 5 ■, 6 □, 7 ◆, 8 ◇) is plotted according to the two objectives. The stars represent an experimentally verified standard refolding condition for each protein.^{27,28} The bisecting line denotes 100% refolding yield and, therefore, the best refolding buffers at different activities. The optimization progress from the start (gen 1) to the end (gen 4, 7, or 8) is highlighted for several gens by the black dashed lines. The lines correspond to the Pareto fronts, the best refolding conditions obtained up to and including the current gen. Consequently, if no better conditions exist the Pareto front remains unchanged and the optimization is not progressing. For the final gen (in brackets: last gen with changes), the experimental error of the Pareto front is highlighted with grey shading. (D) The second, independent optimization approach (modified by constraining redox conditions) of LYZ.

For GLR, the best enzymatic activities (100–120 U mg⁻¹) of both optimization approaches were detected in refolding conditions with ~95% refolding yield. This is considerably higher than the reference refolding condition, which only achieved 61% refolding yield. The original work²⁷ reported up to 70% refolding (depending on protein concentration and renaturation time), which corresponds nicely to the experimentally observed $61 \pm 18\%$. However, improved yields and underlying activities were observed in conditions that are strikingly different to this reference. Our best refolding buffers were (a) far more complex (addition of various additives) and (b) featured an alkaline pH of 8.5 (details in Table III).

GLK refolded very easily with maximum yield, even in the buffer of the functional assay (50 mM HEPES, pH 7.5, 150 mM KCl, 10 mM MgCl₂), that was included as a control. However during the optimization, we obtained far higher activities of up to 266 ± 13 U mg⁻¹ (Table III).

For the fourth enzyme (LYZ), the best result of the second optimization approach (74%) was slightly worse than the reference (82%). However, the underlying activities were ~40% higher. The two optimum refolding conditions (LYZ A/B, Table III) are very similar, differing only in the pH value. They are strikingly different from the reference condition.²⁸ A further, highly interesting result was the activity of native LYZ in different conditions. Our best

Table III. The Best Identified Refolding Conditions (A and B) and Literature Refolding Controls (R) for GFP, GLR, GLK, and LYZ

Refolding condition	Nat	Ref	Yield
GFP A: 40 mM phosphate buffer, pH 7.0, 100 mM NaCl, 10 % v/v glycerol, 50 mM arginine, 50 mM glutamine, 5 mM EDTA, 7.5 mM DTT	18,100 ± 5110	19,862 ± 1743	100 ± 37
GFP B: 50 mM Tris-HCl, pH 7.0, 250 mM NaCl, 15 % v/v glycerol, 100 mM arginine, 50 mM glutamine, 2.5 mM TCEP	25,700 ± 7500	24,885 ± 5717	97 ± 51
GFP R: ²⁶ 40 mM phosphate buffer, pH 7.5, 300 mM NaCl, 50 mM arginine, 50 mM glutamine, 5 mM DTT	28,480 ± 3330	24,900 ± 4152	87 ± 9
GLR A: 100 mM MOPS, pH 8.5, 150 mM NaCl, 20 mM KCl, 500 mM arginine, 50 mM glutamine, 5 mM EDTA, 0.06 mM Tween, 2.5 mM DTT	101 ± 5	96 ± 8	95 ± 12
GLR B: 50 mM MOPS, pH 8.5, 300 mM NaCl, 0.1 w/v % PEG-4000, 100 mM arginine, 100 mM glycine, 2 mM EDTA	120 ± 8	112 ± 18	94 ± 20
GLR R: ²⁷ 20 mM phosphate buffer, pH 6.9, 0.5 mM EDTA, 2 mM DTT	95 ± 20	60 ± 15	61 ± 18
GLK A: 20 mM HEPES, pH 9.5, 350 mM NaCl, 0.05 w/v % PEG-4000, 5 mM EDTA, 5 mM DTT	213 ± 23	236 ± 8	100 ± 14
GLK B: 20 mM Tris-HCl, pH 9.5, 50 mM NaCl, 0.15 w/v % PEG-4000, 50 mM arginine, 50 mM glutamine, 5 mM EDTA, 7.5 mM DTT	266 ± 13	236 ± 8	89 ± 7
GLK R (assay buffer): 50 mM HEPES, pH 7.5, 150 mM KCl, 10 mM MgCl ₂	137 ± 29	198 ± 24	100 ± 28
LYZ A: 100 mM Tris-HCl, pH 7.5, 0.05 w/v % PEG-4000, 50 mM arginine, 100 mM glutamine, 25 mM glycine, 1 mM GSSG	16,8 ± 1,3	12,4 ± 0,9	74 ± 11
LYZ B: 100 mM Tris-HCl, pH 7.0, 0.05 w/v % PEG-4000, 50 mM arginine, 100 mM glutamine, 25 mM glycine, 1 mM GSSG	16,3 ± 1,5	10,2 ± 1,2	62 ± Na
LYZ R: ²⁸ 50 mM Tris-HCl, pH 8.0, 1 mM EDTA, 0.5 mM GSH, 5 mM GSSG	11,2 ± 1,8	9,2 ± 1,8	82 ± 16

Nat, native activity; Ref, refolded activity (GFP fluorescence intensity at 408 nm, GLR and GLK specific activity in U mg⁻¹, LYZ activity according to the EnzChek[®] assay in s⁻¹); yield, relative refolding yield in %.

Listed are the composition, the individual activities of the native and refolded protein, and the yield achieved in the respective refolding conditions.

refolding conditions (LYZ A and B) and the reference (LYZ R) showed very high activities (11–17 s⁻¹). In comparison, the assay buffer (0.1M sodium phosphate, 0.1M NaCl, pH 7.5, and 2 mM sodium azide) performed up to threefold worse (6.0 ± 1.4 s⁻¹). Considering equal protein concentration (5 µg mL⁻¹) and specific activity (70,000 U mg⁻¹), our refolding buffers would consequently achieve up to threefold higher specific activities.

Discussion

This work introduces a new rational strategy for the refolding of proteins from inclusion bodies. It combines screening and optimization of refolding yields in one process with the help of a GA. The primary goal was to design a simple experimental method allowing the optimization of a variety of different proteins with the same routine procedure. The general applicability was proven with the successful refolding of four structurally and functionally different model proteins. The applied algorithm strength pareto algorithm 2 (SPEA 2)^{29,30} is able to optimize several objectives in parallel. In the first two approaches, we successfully optimized both refolding yield and experimental costs for GFP and GLR. The cost optimization was clearly progressing toward the

minimum (Fig. 2). However, even for GFP and GLR, the reference is less expensive than the illustrated conditions (Table III) demonstrating that the cost optimization with the algorithm was not saturated. This indicates that a better cost performance could have been accomplished by continuing the optimization in this direction.

The first optimization approaches resulted in many different refolding conditions with 100% relative yield. Consequently, the algorithm was limited and could only optimize the costs. An additional attribute was the underlying activity, which varied between 40 and 100 U mg⁻¹ (GLR) for refolding conditions with 100% yield. For the following approaches, we optimized the respective native and refolded activities to obtain buffers with highly active protein and maximum refolding yield. In doing so, we optimized the refolding of GLR, GLK, and LYZ. The maximum refolding yields are comparable or better (GLR) than previously observed.^{26–28} Even for the disulfide-bridged LYZ, conditions with similar refolding yields (74%) as described earlier were achieved. Furthermore, the respective activities of the native and refolded protein achieved here were higher in all cases investigated. This is especially interesting regarding the use in enzymatic

assays or in the context of industrial biotransformations. For these purposes, high activities are desired and the optimized activities after refolding might allow the concomitant detection of optimized storage, formulation, and assay buffers.

Interestingly, LYZ was far more difficult to optimize. Of course, we were aware that oxidizing conditions are necessary to efficiently refold LYZ and that our algorithm normally favors reducing conditions. Nevertheless, we were interested to see if the algorithm allows recognizing such a bias in the conditions. Surprisingly, the trend toward oxidizing conditions was clearly visible. This enables an educated user to recognize an unknown, potentially disulfide-bridged protein and to adjust the parameter space. After modifying the experimental design, a second optimization approach performed quite well. Although 80% of the experiments in this approach showed no refolding activity, the analysis of the respective dataset allowed the recognition of a trend. This high negative background seems to be tied to the ionic strength (data not shown). As also reported previously, LYZ activity is drastically reduced at higher ionic strengths.³¹ Thus, a further adjusting of the dataset might allow achieving even higher activities.

This example (LYZ) demonstrated two critical issues: (a) the successful refolding of a disulfide bridge containing protein with our approach and (b) the adaptability and flexibility to address critical parameters while optimizing a specific protein. Taken together, it is striking that the algorithm allows detecting and overcoming limitations arising from single parameters.

Regarding the composition of our best refolding buffers, several trends became visible. All three proteins that contain no disulfide bridges (GFP, GLR, and GLK) preferred reducing conditions with either DTT or TCEP, probably because the reducing conditions prevent oxidation of the free SH-groups via air oxygen. Furthermore, typical refolding additives like the amino acids arginine and glutamine as well as glycerol and PEG seemed to have positive effects for all analyzed proteins (Table III), whereby all additives behaved equally and showed no obvious synergistic effects. In contrast, the buffer substance (phosphate, HEPES, MOPS, or Tris-HCl) had little effect on the refolding yields. Negative factors were the tested metal cations (Cu^{2+} , Zn^{2+} , Mg^{2+} , and Mn^{2+}) and one detergent, sodium deoxycholate. Refolding conditions with either of them present showed no success (data not shown).

In summary, the described work provides a new experimental strategy for the challenging task of screening and optimizing protein refolding from inclusion bodies. Our results demonstrate the strength of this new strategy, providing optimized refolding yields and activities for all four model pro-

teins. Furthermore, our entire optimization strategy seems principally suitable for automatization, as previous studies showed that protein refolding by dilution is generally automatable using pipetting robots.^{11,12}

Materials and Methods

Materials

All chemicals were reagent grade and were obtained from Sigma-Aldrich. Proteins were either purchased in purified form (Sigma-Aldrich) or expressed in *E. coli* as described previously.²⁶

Genetic algorithm

The basis for the optimization strategy was the GA SPEA 2. Fitness assignment, selection, and clustering techniques used are described in detail in the original work.^{29,30} In principle, every individual of the current gen was assigned a fitness based on its dominance (explained below). After generating a mating pool by tournament selection, evolutionary operators (crossover and mutation) were applied resulting in the next gen.

Critical for the optimization were both the objectives and the fitness assignment during the selection process. SPEA 2 is a multiobjective algorithm, which is able to optimize several objectives in parallel, for example, refolding yield, costs, or activities. Optimizing several objectives results in a set of best solutions, as it is not possible to select a single best one. SPEA 2 fitness assignment is based on the pareto principle of dominated and dominating solutions (Fig. 5). Each experimental condition (point) is evaluated in terms of the two optimization objectives (i.e., efficient refolding, low costs, and high activities). A point with a lower value for either or both objectives is dominated. All points that are not dominated by others constitute the “best” experiments, their sum is termed pareto front.

SPEA 2 was implemented in MATLAB (Mathworks, R2009a), and a user friendly Excel (Microsoft, 2003) based file exchange was established. SPEA 2 was used with the following optimization parameters: population size 22 (defined by parameter space), crossover points two, mutation rate one percent per bit, and other parameters were left default.³⁰ The number of gens and thus experiments were varied according to the experiments and depended on the optimization progress of the respective protein. The experimental design space with all tested substances and upper and lower limits is described in detail in the Results section. The problem was coded in bit form with a length (L) of the binary string of 32. Considering the number of refolding conditions ($M = 22$) tested in every generation, it is possible to calculate the probability to reach

Figure 5. Exemplary diagram of a principle data set obtained for each gen. Inefficient solutions (white) are dominated by efficient ones (black), these have greater values in both objectives 1 and 2. Summing up all efficient solutions yields the optimal pareto front (black-dashed line). Grey shading represents the error rates of the data points of the pareto front.

each point in the search space via crossover using the following equation²⁰: $P = (1 - 0.5^{M-1})^L \geq 99.99\%$. P indicates that if the chosen population size is adequate for the complexity of the search space with the given setup, it would even be possible to expand the design space, for example, testing more substances or concentrations in future optimizations.

Screening of protein refolding in solution

New gen was calculated in MATLAB and exported into an excel sheet, which contained the respective 22 refolding conditions of the current gen. After that experimental evaluation results were stored in Excel and imported to MATLAB to calculate the next gen.

Proteins were denatured in presence of guanidine hydrochloride (Gdn-HCl) at room temperature. Specifically, GFP and GLK were incubated overnight in 50 mM Tris-HCl, pH 7.5 with 6M Gdn-HCl (GFP) or 50 mM Tris-HCl, pH 8.0 with 6M Gdn-HCl and 5 mM DTT (GLK).²⁶ GLR and LYZ were incubated for 3 h in 100 mM phosphate buffer, pH 6.9 with 5 mM DTT and 6M Gdn-HCl.²⁸ Denaturation was verified via the respective activity assay. Each gen of refolding experiments consisted of 22 different refolding conditions calculated by the GA. For folding, the denatured protein was rapidly diluted to 50–200 folds to a final protein concentration of 5 $\mu\text{g mL}^{-1}$ in the respective refolding buffer. For example, 10 μL of denatured LYZ (1000 $\mu\text{g mL}^{-1}$) were added to 1990 μL refolding buffer and mixed. For the native control, nondenatured protein with the same concentration was equally diluted. The samples were incubated overnight at 10°C in 2.2 mL, 96-well plates (Sarstedt, Germany).

Refolding success was determined via functional assays. To negate buffer effects, both negative (only buffer) and native controls (with nondenatured protein) were deemed necessary. Thus, for each of the 22 conditions one negative control, three native, and four refolded samples were analyzed. The indicated refolding yields represent the average activity, given as relative percentage of the respective native control.

As certain refolding additives, for example, arginine and redox agents are expensive compounds, the overall cost of the refolding buffer was considered. Using the pricing of the provider (Sigma-Aldrich), individual costs of the respective compounds were summarized and indicated as overall costs of the respective refolding buffer (€ mL^{-1}).

Determination of enzyme activities

The structural integrity of GFP (variant F64L and S65T)³² was determined as described in Ref. 26. GLK ATPase-activity was measured in an ATP-regenerating system coupled to β -Nicotinamide adenine dinucleotide, reduced (NADH)-consumption, which was monitored at 340 nm.³³ The measurement was carried out in the presence of 2.5 mM Adenosine 5'-triphosphate (ATP) and 800 μM D-glucose. LYZ activity was analyzed with the EnzChek[®] Lysozyme Assay Kit (Molecular Probes) with one minor modification. Instead of performing endpoint measurements, we determined the reaction kinetics at 37°C. GLR was determined according to Ref. 34 by measuring the decrease of the substrate β -Nicotinamide adenine dinucleotide phosphate, reduced (NADPH) at 340 nm for 15 min. The assay was adjusted to 96-well plates and 250 μL total volume. The assay buffer contained 75 mM phosphate buffer, pH 7.6, 2.6 mM EDTA, 1 mM GSSG, and 0.09 mM NADPH.

Acknowledgments

The authors thank the administrators of REFOLD for their data and support, Elena Kunold for practical assistance, and Tetyana Dashivets for GFP and GLK.

References

1. Rudolph R, Lilie H (1996) In vitro folding of inclusion body proteins. *FASEB J* 10:49–56.
2. Middelberg AP (2002) Preparative protein refolding. *Trends Biotechnol* 20:437–443.
3. Jungbauer A, Kaar W (2007) Current status of technical protein refolding. *J Biotechnol* 128:587–596.
4. Clark E (2001) Protein refolding for industrial processes. *Curr Opin Biotechnol* 12:202–207.
5. Burgess R (2009) Refolding solubilized inclusion body proteins. *Methods Enzymol* 463:259–282.
6. Cabrita LD, Bottomley SP (2004) Protein expression and refolding—a practical guide to getting the most out of inclusion bodies. *Biotechnol Annu Rev* 10:31–50.

7. Cowan RH, Davies RA, Pinheiro TT (2008) A screening system for the identification of refolding conditions for a model protein kinase, p38alpha. *Anal Biochem* 376: 25–38.
8. Lin L, Seehra J, Stahl ML (2006) High-throughput identification of refolding conditions for LXRbeta without a functional assay. *Protein Expr Purif* 47: 355–366.
9. Willis MS, Hogan JK, Prabhakar P, Liu X, Tsai K, Wei Y, Fox T (2005) Investigation of protein refolding using a fractional factorial screen: a study of reagent effects and interactions. *Protein Sci* 14:1818–1826.
10. Trésaugues L, Collinet B, Minard P, Henckes G, Aufrère R, Blondeau K, Liger D, Zhou C, Janin J, Van Tilbeurgh H, Quevillon-Cheruel S (2004) Refolding strategies from inclusion bodies in a structural genomics project. *J Struct Funct Genomics* 5:195–204.
11. Scheich C, Niesen FH, Seckler R, Büsow K (2004) An automated in vitro protein folding screen applied to a human dynactin subunit. *Protein Sci* 13:370–380.
12. Vincentelli R, Canaan S, Campanacci V, Valencia C, Maurin D, Frassinetti F, Scappucini-Calvo L, Bourne Y, Cambillau C, Bignon C (2004) High-throughput automated refolding screening of inclusion bodies. *Protein Sci* 13:2782–2792.
13. Tobbell DA, Middleton BJ, Raines S, Needham MR, Taylor IW, Beveridge JY, Abbott WM (2002) Identification of in vitro folding conditions for procathepsin S and cathepsin S using fractional factorial screens. *Protein Expr Purif* 24:242–254.
14. Armstrong N, de Lencastre A, Gouaux E (1999) A new protein folding screen: application to the ligand binding domains of a glutamate and kainate receptor and to lysozyme and carbonic anhydrase. *Protein Sci* 8:1475–1483.
15. Rechenberg I (1973) *Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Stuttgart: Friedrich Frommann Verlag.
16. Holland H (1975) *Adaption in natural and artificial systems, an introductory analysis with application to biology, control and artificial intelligence*. Ann Arbor, MI: The University of Michigan Press.
17. Back T, Hammel U, Schwefel H (1997) Evolutionary computation: comments on the history and current state. *IEEE Trans Evol Comput* 1:3–17.
18. Gobin OC, Joaristi AM, Schüth F (2007) Multi-objective optimization in combinatorial chemistry applied to the selective catalytic reduction of NO with C₃H₆. *J Catal* 252:205–214.
19. Gobin OC, Schüth F (2008) On the suitability of different representations of solid catalysts for combinatorial library design by genetic algorithms. *J Comb Chem* 10: 835–846.
20. Weuster-Botz D (2000) Experimental design for fermentation media development: statistical design or global random search? *J Biosci Bioeng* 90:473–483.
21. Havel J, Link H, Hofinger M, Franco-Lara E, Weuster-Botz D (2006) Comparison of genetic algorithms for experimental multi-objective optimization on the example of medium design for cyanobacteria. *Biotechnol J* 1:549–555.
22. Link H, Weuster-Botz D (2006) Genetic algorithm for multi-objective experimental optimization. *Bioprocess Biosyst Eng* 29:385–390.
23. Amin AA, Bottomley SP, Chow MK, Fulton KF, Whisstock JC, Buckle AM (2006) REFOLD: an analytical database of protein refolding methods. *Protein Expr Purif* 46:166–171.
24. Buckle AM, Devlin GL, Jodun RA, Fulton KF, Faux N, Whisstock JC, Bottomley SP (2005) The matrix refolded. *Nat Methods* 2:3.
25. Qoronfleh MW, Hesterberg LK, Seefeldt MB (2007) Confronting high-throughput protein refolding using high pressure and solution screens. *Protein Expr Purif* 55:209–224.
26. Dashivets T, Wood N, Hergersberg C, Buchner J, Haslbeck M (2009) Rapid matrix-assisted refolding of histidine-tagged proteins. *Chembiochem* 10:869–876.
27. Nordhoff A, Tziatzios C, van Den Broek JA, Schott MK, Kalbitzer HR, Becker K, Schubert D, Schirmer RH (1997) Denaturation and reactivation of dimeric human glutathione reductase—an assay for folding inhibitors. *Eur J Biochem* 245:273–282.
28. Hevehan DL, De Bernardez Clark E (1997) Oxidative renaturation of lysozyme at high concentrations. *Biotechnol Bioeng* 54:221–230.
29. Zitzler E, Laumanns M, Thiele L (2001) SPEA2: improving the strength pareto evolutionary algorithm. TIK Report 103. Zürich: Eidgenössische Technische Hochschule Zürich (ETH).
30. Zitzler E, Laumanns M, Thiele L, SPEA2: improving the strength pareto evolutionary algorithm for multi-objective optimization. In: KC Giannakoglou, DT Tsahalis, J Periaux, KD Papaliliou, T Fogarty, Eds. (2002) *Evolutionary methods for design, optimisation and control*, Barcelona, Spain: CIMNE, pp 95–100.
31. Davies RC, Neuberger A, Wilson BM (1969) The dependence of lysozyme activity on pH and ionic strength. *Biochim Biophys Acta* 178:294–305.
32. Topell S, Hennecke J, Glockshuber R (1999) Circularly permuted variants of the green fluorescent protein. *FEBS Lett* 457:283–289.
33. Nørby JG (1988) Coupled assay of Na⁺,K⁺-ATPase activity. *Methods Enzymol* 156:116–119.
34. Mavis D, Stellwagen E (1968) Purification and subunit structure of glutathione reductase from bakers' yeast. *Biol Chem* 4:809–814.