Available online at www.sciencedirect.com
SCIENCE DIRECT®
Progress in Biophysics and Molecular Biology 88 (2005) 285–309
Progress in Biophysics & Molecular Biology
www.elsevier.com/locate/pbiomolbio

Review

Protein crystallization: virtual screening and optimization

Lawrence J. DeLucas[a,*], David Hamrick[b], Larry Cosenza[b], Lisa Nagy[a], Debbie McCombs[a], Terry Bray[a], Arnon Chait[c], Brad Stoops[c], Alexander Belgovskiy[c], W. William Wilson[d], Marc Parham[e], Nikolai Chernov[f]

[a]*Center for Biophysical Sciences and Engineering, The University of Alabama at Birmingham, Birmingham, AL, USA*
[b]*Diversified Scientific Inc., Birmingham, AL, USA*
[c]*ANALIZA Inc., Bay Village, OH, USA*
[d]*Mississippi State University, Mississippi State, MS, USA*
[e]*Interactive Analysis, Bedford, MA, USA*
[f]*Natural Sciences and Mathematics, The University of Alabama at Birmingham, Birmingham, AL, USA*

Available online 30 September 2004

Abstract

Advances in genomics have yielded entire genetic sequences for a variety of prokaryotic and eukaryotic organisms. This accumulating information has escalated the demands for three-dimensional protein structure determinations. As a result, high-throughput structural genomics has become a major international research focus. This effort has already led to several significant improvements in X-ray crystallographic and nuclear magnetic resonance methodologies. Crystallography is currently the major contributor to three-dimensional protein structure information. However, the production of soluble, purified protein and diffraction-quality crystals are clearly the major roadblocks preventing the realization of high-throughput structure determination.

This paper discusses a novel approach that may improve the efficiency and success rate for protein crystallization. An automated nanodispensing system is used to rapidly prepare crystallization conditions using minimal sample. Proteins are subjected to an incomplete factorial screen (balanced parameter screen), thereby efficiently searching the entire ''crystallization space'' for suitable conditions. The screen conditions and scored experimental results are subsequently analyzed using a neural network algorithm to predict new conditions likely to yield improved crystals. Results based on a small number of proteins suggest that the

*Corresponding author. Fax: +1 205 934 2659
E-mail address: delucas@cbse.uab.edu (L.J. DeLucas).

0079-6107/$ - see front matter
doi:10.1016/j.pbiomolbio.2004.07.008

combination of a balanced incomplete factorial screen and neural network analysis may provide an efficient method for producing diffraction-quality protein crystals.

Keywords: Nanocrystallization; Neural network optimization

Contents

1. Introduction

Structural biology has become increasingly important due to the availability of entire genome sequences for a variety of organisms including human, mouse, dog, plus many different bacteria and viruses (http://www.nigms.nih.gov/funding/psi/lay_summary.html). The collective sequence information alone can suggest associations between specific proteins and biological processes, but unfortunately, novel sequences (minimal or no homology with known protein sequences) provide little insight into a protein's mechanism of action or biological function. Three-dimensional structure information plays a critical role determining a protein's biological function and mechanism of action. In addition, protein structural information can aid in the development of novel medications, vaccines and diagnostics. The growing importance of structural biology in basic and applied biomedical research has resulted in an exponential demand for protein structure information. Governmental science agencies from several countries have initiated large programs to develop high-throughput structural genomic technologies to provide structure determinations for several thousand novel proteins (http://www.nigms.nih.gov/news/meetings/airlie.html#agree). Significant funding for structural genomics currently exists in the US, Canada, the European Union, Israel, China, and Japan. It is believed that experimentally determined structures of novel protein folds will aid computational modeling of related sequence homologs to ultimately yield structural information for the majority of sequenced genes. The international structural genomics

effort has led to several significant improvements in bioinformatics, cloning, protein expression, and purification, methods for preparing crystallization experiments and X-ray data collection (Krupka et al., 2002; Luft et al., 2001; Stevens, 2000).

The US National Institutes of Health (NIH) established a structural genomics program with the goal of "encouraging research on the development of methodology and technology underpinning the emerging field of structural genomics, whose goal is the understanding of protein structural families, structural folds, and the relation of structure and function". One such program (P50-GM62407), "the Southeastern Collaboratory for Structural Genomics (SECSG)", involves researchers from a consortium of universities, including the University of Georgia (UGA, P.I., B.C. Wang), Georgia State University (GSU), the University of Alabama at Birmingham (UAB), the University of Alabama in Huntsville (UAH), and Duke University. The complete genomes from *Caenorhabditis elegans* and *Pyrococcus furiosis*, plus selected genes from the human genome, were chosen for this collaborative program. A total of nine such centers have been created in the US with the majority of the centers selecting prokaryotic organisms in an effort to simplify the entire process, particularly high-throughput protein expression protocols (only three centers have included eukaryotic systems as all or a portion of their targets). A number of techniques have been employed by the centers for determining protein structures, including X-ray crystallography, nuclear magnetic resonance spectroscopy, and mass spectrometry. Of these, X-ray crystallography remains the only method routinely used to determine structures of large biomolecules (i.e. MW in excess of 20,000 Da). Unfortunately, in spite of the fact that most of the target organisms are prokaryotic, the center's success rates for producing structures are extremely low. Table 1 provides a summary of the status for seven of the nine US centers (these represent the original centers that have been in operation for 3 years).

Inspection of Table 1 suggests that there are two major bottlenecks, the production of soluble protein and the production of crystals of suitable diffraction quality to enable a structural solution. This paper addresses the second bottleneck, crystal production.

Recent efforts in structural genomics have produced thousands of new proteins for study in structural biology and drug design projects. The number of new proteins available will continue to increase significantly in the next several years as additional investigators become involved in this important research. However, there are several problems that are currently providing a barrier to efficient, cost-effective crystallization. One involves the production of ample amounts of purified protein. In spite of the sophisticated new methods for enhancing protein production, this remains as a serious impediment to cost-effective high-throughput crystallography (Edwards et al., 2000; Waldo et al., 1999; Goulding and Perry, 2003). Other problems include the ability to find crystallization conditions or optimize initial crystallization conditions sufficiently to provide diffraction-quality crystals.

Table 1
Summary of crystallization results from NIH Structural Genomics Initiative Statistics based on seven NIH Structural Genomics Centers

	Cloned	Soluble proteins	Crystallized	Diffraction-quality	Structures	Deposited in PDB
Total number	21,149	7652	1793	766	555	409
Percentage		36.2	23.4	10.0	7.3	5.3

Yet, with the advent of high-throughput liquid handling and crystallization systems (Krupka et al., 2002; Luft et al., 2001; Stevens, 2000b) it is now relatively easy to prepare a thousand or more crystallization experiments. Thus, one can conclude that the number of conditions tested is not solely responsible for these dismal crystallization success rates. Clearly, other factors must be involved and as a result, alternative strategies must be adopted to adequately address this problem. One novel approach involves crystallization of a homologous protein from another organism (Campbell et al., 1972). Since in many cases, the differences in sequence are minor, and these differences are often found on the surface or in regions without direct biological activity (i.e. not in the active site of an enzyme), coordinate information from the X-ray structure of the homologous protein can be used to accurately model the structure of the original protein. Alternatively, one can modify the target protein's surface by introducing cofactors, or other additives, antibodies, or removal of carbohydrate, all in an effort to produce more suitable crystalline lattice contacts (Davis et al., 1990; Ostermeier et al., 1997; Prongay et al., 1990; DeLucas et al., 2003; Baker et al., 1994; Gruenunger-Leitch et al., 1996; Kostrewa et al., 1997; Oefner et al., 2000). Limited proteolysis can also help provide a protein form that is, by chance, more conducive to crystallization (McPherson, 1982). Introduction of point mutations, truncations, or deletions has also been demonstrated to help improve crystallization success rates (Lawson et al., 1991; McElroy et al., 1992; D'Arcy et al., 1999, Longenecker et al., 2001; Mateja et al., 2002; Charron et al., 2002, Chen et al., 1996; Dale et al., 1994; Ay et al., 1998, Betton et al., 1997; Nagi and Regan, 1997; Nugent et al., 1996; Thompson and Eisenberg, 1999; Zhou et al., 1996).

A critical component of X-ray crystallography is obtaining well-ordered crystals of the target protein. This effort traditionally requires screening thousands of solutions with varying chemical compositions. Several commercially available crystallization kits, although quite helpful (Jancarik and Kim, 1991), are based on previous success analysis. Such kits tend to provide fine screens in areas where other proteins have been successfully crystallized. This approach results in excessive screening in localized regions of "crystallization space" and little or no screening in other regions. This information combined with the fact that crystallization success rates remain extremely low suggests that there is a significant need for more efficient and effective methods for determining protein structures.

2. Strategy

Crystallization screens that efficiently search the entire "crystallization space" should screen each variable plus all combinations of these variables at appropriate increments to avoid wasting material on experiments that are likely to yield equivalent results and to ensure that all crystallization possibilities have been sampled. "Hits" obtained in the initial screening process are used to design an optimization screen to further improve the initial crystallization results. Typically, this is accomplished by performing a fine screen centered around the best initial "hit". However, results from a statistically balanced screen provide a great deal of information regarding the importance of each variable as well as combinations of two or more variables versus outcomes (scores). It is not possible for a trained crystallographer to assimilate all of this information to accurately and efficiently design new conditions that optimize the levels (i.e. component

concentrations, temperature, etc.) of specific variables and their combinations. Predictive algorithms (neural networks) are known to be particularly useful for analyzing complex relationships between a large number of variables with respect to the results these combinations produce. If sufficient information is provided (i.e. a range of results from clear drops to crystals for the screening conditions), it may be possible for a neural network to train itself to predict new outcomes based on the different combinations of these variables. That is, based on initial screen conditions and the results produced by the screen, the neural network could perform a virtual screen of all possible combinations and levels of variables to predict optimum conditions likely to yield larger and higher-quality crystals.

In an effort to minimize the total amount of protein required to screen for suitable crystallization conditions, the Center for Biophysical Sciences and Engineering (CBSE), ANALIZA Inc., and Diversified Scientific Inc., have explored three different complementary approaches; an incomplete factorial screen, a high-throughput nanoliter crystallization robot, and a neural network software program capable of using initial screen results to predict future conditions that are likely to yield crystals. The incomplete factorial screen allows a small number of experiments to be performed that sample all possible experiments in a statistically robust manner. This approach should allow for efficient determination of solution conditions suitable for crystallizing proteins by performing experiments that take into account the independent and interdependent influences of each experimental parameter.

3. Experimental investigations

To enable a comprehensive search in a large parameter space for optimal crystallization conditions using sub-milligram protein quantities, a modular line of high-throughput crystallization and inspection workstations has been developed. The process begins with a statistically-based screen optimization that directs the production of specialized libraries of crystallization conditions. The libraries and the proteins are subsequently combined using a novel nanoliter-range crystallization screening system (*NanoScreen*™). An automated intelligent high-resolution inspection system (*CrystalScore*™) is then deployed to periodically examine and classify the optimal starting conditions for the subsequent scale-up crystallization experiments. This system has recently been upgraded to include a neural network crystallization prediction program.

The high-throughput nanoliter crystallization robot significantly reduces the scale of each experiment, allowing the use of as little as 0.6 μg of protein per screening experiment condition. The use of neural network software programs facilitates prediction of probable crystallization conditions based on results from a small number of experiments. This can further improve the efficiency of protein crystallization screening experiments by learning from prior experimental results and predicting new conditions that should produce crystals. The three technologies work in tandem to facilitate highly efficient and effective screening of protein crystallization conditions. An automated system that combines all three approaches was recently developed in an effort to produce a more efficient and successful method for macromolecular crystallization. The following describes our initial experimental investigations.

3.1. Incomplete factorial screen

The following provides a summary of the rationale used to develop the incomplete factorial screen (a more complete description is provided in DeLucas et al. 2003). The conditions chosen to crystallize a macromolecule exploit and control a variety of energetic differences. The pH, for example, determines the charge on the molecule, directly influencing the energetics of its interaction with the bulk solvent. The addition of counter ions shields surface charges and changes the chemical potential of the solvent. Polymeric alcohols sequester water away from the macromolecule and may interact with it as well. Certain other components, such as divalent cations and metals, may interact directly with the macromolecules and moderate lattice contacts. Although some macromolecules are crystallized solely by temperature change, in most cases the buffers that stabilize the solution and crystalline states have different compositions. This is often accomplished either by evaporation or dialysis. The kinetics of the growth process is affected by the crystallization method used, such as vapor diffusion, dialysis, or controlled evaporation.

A preliminary crystallization screen searches broadly though "crystallization space", a multidimensional set of possible components, concentrations, and physical conditions. The goal is good outcomes (i.e. leads to diffraction-quality crystals). The experiment number and type for the initial screen are chosen to reflect a reasonable balance between thoroughness and cost; and a reasonable expectation of where the desired results may be found. For example, excessive trials will return many repeated results from similar experiments, wasting material, time, and manpower. An insufficient number of experiments may miss a region that would have ultimately led to crystals. It is also important to avoid time- and sample-consuming general trials in areas of chemical space that would rarely, if ever, produce crystals. These areas might be accessed later by widening the search if initial trials fail to produce leads.

There are many strategies available to search for crystallization conditions. Commercial screens use sparse matrix methods, in which the experiments are clustered around conditions that have already given crystals in the past. The advantage of this approach is that when a protein is crystallized under one set of conditions, it will often exhibit "hits" in other conditions as well. The disadvantage is that some areas of "crystallization space" are neglected. Random screens (such as CRYSTOOL) sample these areas, but in both cases it is difficult to glean information from the collected results because the screens are not balanced.

With the assistance of Professor Charles Carter (University of North Carolina, Chapel-Hill), CBSE scientists have been developing a set of conditions to construct an efficient screen. A general model is used that incorporates factors whose levels are mathematically balanced. Each possible level of a factor is sampled an equal number of times. In our 360-experiment screen, each of the six anionic precipitants is sampled 60 times. Binary combinations are also balanced. Third order and higher combinations are distributed randomly. A statistical computer program INFAC (Carter, private communication) is used to construct the balanced matrices that encode the experiments. A spreadsheet program translates the matrices into chemical recipes for our solution handling robots to construct. This balanced design facilitates the determination of which factor levels are most suitable for crystallization. For example, a comparison of the average score of all the experiments that contain chloride as an anionic precipitant to the overall average (and the other anions) allows one to determine whether chloride is the best choice for the anion. Some binary combinations will show obvious synergy, such as the combination of pH and anion choice.

Anions show varying effectiveness as precipitating agents as the net charge on the protein changes (Hofmeister, 1888). The temperature can have a significant effect on the solubility of the macromolecule, also affecting the solution properties for other components. Several buffers have a temperature dependence on their pKa's, a characteristic that may be exploited. The pH affects the net charge on the molecule and the charge state of the surface amino acids. In a screen, the precipitant concentration is varied relative to the macromolecule concentration. This avoids areas of crystallization space that would never give crystals. Both organic precipitants, such as polymeric alcohols (polyethylene glycols), and ionic salts are useful as precipitants. Many ionic salts are used as precipitants, and have particular charges, tendencies to exhibit polar effects, sizes and solution activities. Many have significant buffering capabilities as well. Using a combination of organic and inorganic precipitants can balance their various properties. Glycerol can stabilize the macromolecule in solution by specific interaction with the surface. The addition of glycerol can shift the nucleation point of the macromolecule independent of other factors. Divalent cations stabilize specific interactions between macromolecule monomers, often increasing the order of a crystal. Many proteins bind specific divalent metals, and often the metals have an anomalous signal that can be used for phasing. Additives such as detergents and arginine can affect the macromolecule conformation, strengthen specific contacts, or reduce non-specific contacts.

Once a screen design has been devised and constructed, the experiments are conducted and the outcomes are scored. A scale was used that includes scores for various crystalline, quasi- and non-crystalline results. A matrix to encode the variables for the screen was calculated using the program INFAC. A screen size was chosen for 360 experiments with 10 variables at six or three levels each. The most balanced matrix was chosen from the 10,000 seeds tested. Screen variables were chosen according to previously published work (Carter and Carter, 1979).

The matrix was translated into a set of recipes using a spreadsheet program. The stock components were mixed using the *RecipeMaker*^{TM} robot, a custom-configured Hamilton ML 4000 liquid handling robot capable of mixing up to 48 different stocks and water. Recipes were prepared in 2 ml block plates and reformatted into standard 384-well plates. For comparison, the proteins were also subjected to a group of commercial screens consisting of Hampton Crystal Screen I and II, MembFac, Natrix, and Emerald Wizard I and II (290 experiments). Comparison between incomplete and sparse matrix screen's ability to crystallize protein suggests that the incomplete factorial method can find a larger number of conditions that are useful for growing crystals. A representative set of 18 proteins (seven from *C. elegans,* three from *P. furiosis*, three proprietary commercial, two commercially available, and three other), were subjected to crystallization screening using both incomplete factorial and sparse matrix models. For this limited sample size, the incomplete factorial model of screening generally resulted in more conditions per protein favoring the growth of crystals. Interestingly, the incomplete factorial model resulted in identifying conditions that appear disparate compared to the results from the sparse matrix screens.

Table 2 summarizes the screening results of 18 proteins for the incomplete factorial screen vs. commercial screens. "Hits" represent those screening conditions that yielded crystalline material. The data suggest that the incomplete factorial screen provides a higher "hit rate" for crystallization conditions than do available commercial screens. In addition, in some cases, crystallization outcomes may only be achievable via the incomplete factorial approach.

Table 2
Comparison of results (crystalline material) obtained using incomplete factorial screen versus available commercial screens

Protein	Incomplete factorial screen hits	Commercial screen hits
C. elegans proteins		
2G1.1	3	5
9C9	5	2
11059b	8	11
18B5	20	17
3D17	2	2
25D10	25	3
74D6	4	1
P. furiosis proteins		
UGA 214	6	1
UGA 220	7	0
UGA 222	2	0
Proprietary commercial proteins		
PX1	32	17
PA1	3	2
TSP	4	3
Commercially available proteins		
Catalase	107	43
α-chymotrypsinogen (ACTP)	50	25
Other proteins		
Variable chitin binding protein (VCBP)	36	8
Collagen-binding protein ACE 40	3	1
Bacterial hyuralonidase	1	2

3.2. High-throughput nanoliter protein crystallization

CBSE scientists recognized several years ago that the demand for a significant quantity of protein for crystal screening would inhibit the ability of researchers to determine structures for a significant percentage of the targeted proteins. In 1997, the CBSE began developing new technologies that would support automated, reduced-scale crystallization screening. Indeed, several companies, academic laboratories, and government laboratories are now actively pursuing reduced-scale crystallization, both with internally developed technologies and commercially available systems (DeLucas et al. 2003; Hosfield et al., 2003; DeTitta et al. 2001; Luft et al., 2001; Mueller et al., 2001; Krupka et al., 2002; Santarsiero et al., 2002). This approach has proven to be an enabling methodology for high-throughput crystallization screening and structure determination.

Macromolecular crystallization is characterized by extremely slow and highly anisotropic molecular attachment kinetics. Large molecules require a longer time (as compared with small molecules) to assemble into a highly ordered crystalline lattice. Critical variables include the solvent structure, the presence of various crystallization agents that modify solvent structure, and

the rate at which protein molecules are transported to the crystalline surface. In large-scale crystal growth it is possible to handle highly viscous solutions, mix, and prepare screening solutions containing minor ingredients in very low proportions. However, reduced-scale crystallization places constraints on these three requirements.

In an effort to simplify the overall requirements of the nanoliter dispensing system, the preparation of the incomplete factorial screen (crystallization solutions) is performed independently and prior to the preparation of the nanocrystallization experiments. Libraries of crystallization solutions are prepared according to previously described statistical design methodologies, using conventional liquid handlers in large volumes (ca. 1 ml). Viscous bulk ingredients can be dispensed at proper proportions and completely mixed in advance of the actual screening experiment. Different screening libraries specific for each class of proteins are prepared and stored in sealed 96 deep-well plates for further reformatting and subsequent use. The actual experiment is reduced to aspirating/dispensing of the screening and protein solutions for the specific crystallization technique chosen.

There are three fundamental approaches to deliver sub-microliter scale volumes. The simplest approach is the spotting method, whereby a pin with an extremely fine point is dipped into a solution and removed. A small amount of solution clings to the tip of the pin and can be transferred to a microarray or slide by touching the pin to the target surface. This method is easy to implement but it is difficult to control the accuracy and precision of the delivered volume. A second approach involves using chambers of defined dimensions to control the volumes delivered. A common application of this method is in microfluidic devices, which utilize fabrication techniques from the computer chip industry to create channels and chambers in silicon that can contain nanoliter volumes. However, controlling fluid movement in this format has proved difficult primarily due to the limitations of valve fabrication. Effective valve fabrication technology has been achieved in soft polymeric material devices (Fluidigm Corporation South San Francisco, CA), but this represents a relatively recent accomplishment. A third method for delivering low volumes involves dispensing technologies. At the outset, the CBSE and ANALIZA determined that the use of active dispensing technologies was most feasible for successfully performing high-throughput nanoliter crystallization. Fundamental technologies capable of dispensing nanoliter volumes were available in the inkjet printing industry that could be adapted to our application. Other approaches were judged to have limitations not easily overcome in a reasonable timeframe. Although the active dispensing approach was judged to be the most feasible, there remained several hurdles that had to be addressed. One of the most challenging problems encountered was in delivery of nanoliter volumes of solutions with widely varying physicochemical properties. This problem was solved by the choice of active dispensing method and by diluting solutions to achieve a more water-like activity. These two approaches resulted in a system capable of dispensing solutions exhibiting a wide range of physicochemical properties with high precision and accuracy (*NanoScreen*™). Other practical problems included dispensing multiple nanoliter volumes simultaneously with similar precision and accuracy, eliminating cross-contamination between solutions, and controlling unintended water loss during experiment preparation. Dispensing of multiple nanoliter solutions was accomplished using a hybrid microfluidic valve (Innovadyne, Santa Rosa, CA) that controlled the fluid flow at each dispensing tip. Cross-contamination was eliminated by use of custom wash stations that rinsed the dispensing tips thoroughly before sets of new solution conditions were aspirated. Water loss during

experiment preparation was virtually eliminated via an automated oil (microbatch) dispense system or by using a humidity chamber (vapor diffusion) to retard evaporation.

Dispensing of sufficiently viscous screening solutions in very small volumes (e.g., low nanoliter quantities) is a vexing problem that rapidly reduces the choice of dispensing technologies. Techniques relying on surface wave excitation and instability, such as piezo capillaries, are inherently limited since they typically rely on small variations in surface tension and viscosity to achieve quantitative accuracy. Screening solutions vary widely in all of their physicochemical parameters, thus requiring a robust dispensing technique to achieve reproducible and quantitative screens. Fast solenoid (a.k.a. drop-on-demand) techniques were chosen because they rely on fluid inertia for dispensing and typically give accurate results across a large operating range.

Low volume screening experiments can be executed in multiple ways. Our initial work concentrated on microbatch (under-oil) crystallization (Chayen, 1997), in which the screening and protein droplet is quickly covered with a mixture of water-impermeable and water-permeable oils, offering the desired kinetic profile (e.g., 1 or 2 weeks for complete drying). Crystallization under oil is particularly suitable for low volume screening in which crystal recovery is not required. Also, rapid evaporation of prepared droplets is a serious issue in low volumes and typical techniques such as base plate cooling and humidity control are not optimal since every droplet (screen) has different colligative properties. For example, base plate cooling below the dew point of water can result in droplets continuing to dry while others, on the same multi-well plate, actually experience hydration. Rapid sealing of individual wells requires sophisticated automation, a level of complexity that was omitted from initial prototypes. Other crystallization techniques such as sitting drops are easily accommodated, and new miniaturized crystallization plates facilitate such experiments. It should be noted that a thin layer of silicon oil can be used to slow the rapid initial evaporation, thereby allowing an entire plate to be prepared before covering the individual experiment chambers.

As discussed before, a preferred method for conducting reduced-scale crystallization screening is through the use of technologies, similar to those employed in ink jet printing devices, which allow small volumes to be dispensed with high accuracy and precision. The CBSE, partnered with ANALIZA Inc. to develop *NanoScreen*™, an automated crystallization system that can dispense 20 nl of solution with $\pm 10\%$ accuracy. This system can prepare more than 1000 crystallization experiments consuming less than 600 μg of protein (assuming a concentration of 10 μg/ml). A wide range of solution viscosities can be accommodated while maintaining accurate dispensing. Droplets ranging in volume from 20 to 200 nl consisting of up to 10% PEG 8000 can be rapidly dispensed with less than 10% error. Higher PEG concentrations are achieved by dispensing 4-fold diluted PEG solutions (making the solutions more water like with similar chemical contents). The microbatch method can be adjusted so that water is continually drawn from the diluted crystallization droplet until the desired %-PEG is obtained. This dilution effect is accounted for in our statistical experiment design strategy.

The *NanoScreen*™ system (Fig. 1) is comprised of several key components. A microfluidic dispensing head is used to handle the protein and crystallization (recipe) solutions. The first-generation system has 12 tips for fluid handling, with two tips dedicated to dispensing protein solutions and the remaining 10 used to deliver the recipe solutions. Motorized stages allow for accurate x; y; and z-axis movement of the head to the various locations on the deck of the *NanoScreen*™ system. The system can perform both microbatch and traditional vapor diffusion

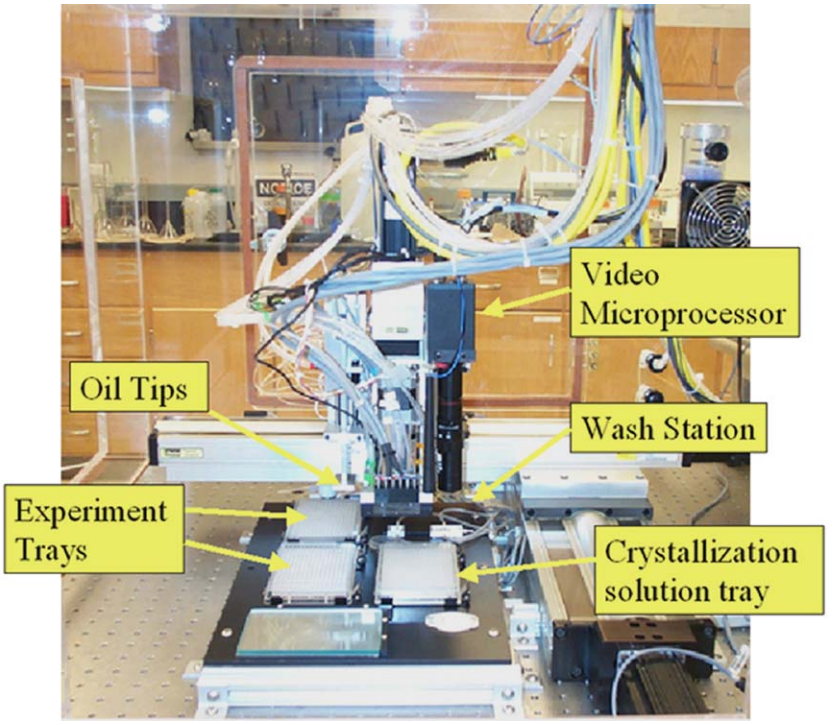

Fig. 1. *NanoScreen*™ crystallization system

(sitting drop) experiments. For microbatch experiments, an oil dispensing subsystem is used to cover the experiment solutions after they are deployed. Wash stations for protein and recipe prevent contamination between solutions and a quality control capability used to calibrate drops and to monitor the dispense mechanism. Custom software controls all aspects of the *NanoScreen*™ operation, including solution aspiration, solution dispensing, tip washing, oil dispensing, movement of the stages, quality control operations, and drop dispensing calibration. The system accommodates experiments in conventional 384-well plates, as well as the new Corning 192 experiment vapor diffusion plate.

The *NanoScreen*™ system operates by aspirating 20 μl of protein solution into two protein tips and 20 μl of 10 different recipe solutions into the 10 recipe tips. Before the first experiments are prepared, drops are dispensed from each tip onto the quality control area to ensure that the tips are dispensing properly. The dispensing head then moves to the first set of experiment wells and dispenses the protein solution. The 10 recipe tips are then automatically positioned over the protein solutions in the experiment wells and recipe solutions from each of these tips are dispensed simultaneously. For microbatch experiments, an oil mixture is immediately dispensed over the experiment solutions to prevent unintended water loss. For vapor diffusion experiments, the experiments are prepared within a constant humidity chamber to minimize unintended evaporation, but coverage with permeable oil also offers an option to arrest the initial rapid evaporation while not interfering with the slow approach to supersaturation. With each set

of 10 experiments prepared, the recipe tips are washed, followed by aspiration of the next set of 10 different recipe solutions for subsequent experiments. This process is repeated until all experiments on a tray have been prepared. The experiment tray is then manually sealed with a Crystal Clear Sealing Tape (Corning 6575) and placed in a constant temperature incubator. The *NanoScreen*™ system has a throughput of 400 experiments/hr. While approximately 40 μl of protein solution is needed as a total working volume, very little of this initial solution is actually consumed. Most of the aspirated protein solution is returned, after completing preparation of all experiments, where it can be recovered and used for future experiments. It should be noted that the first generation *NanoScreen*™ system is a prototype, with improvements already under development to increase throughput by a factor of 10.

3.3. Experiment imaging

Images are acquired and stored using the *CrystalScore*™ imaging system developed by Diversified Scientific Inc. (Fig. 2). The *CrystalScore*™ system allows for automated image acquisition of each experiment, archiving of sequential images, data storage, and automatic determination of crystal location, size, and number for any experiment. An automated imaging platform, consisting of a robotic arm, a custom *CrystalScore*™ system, and incubators housing the experiment trays on a carousel is integrated with a bar code reader, plate imaging scheduler, central server for housing image databases, and client computers for accessing experiment databases and evaluating crystallization results.

Other *CrystalScore*™ capabilities include filtering and sorting relational databases for ''hits'' and growth trends, storage of 2 M database records on local workstations, remote database connectivity for Oracle 9I, IBM DB2, and Microsoft SQL servers, database report generation to

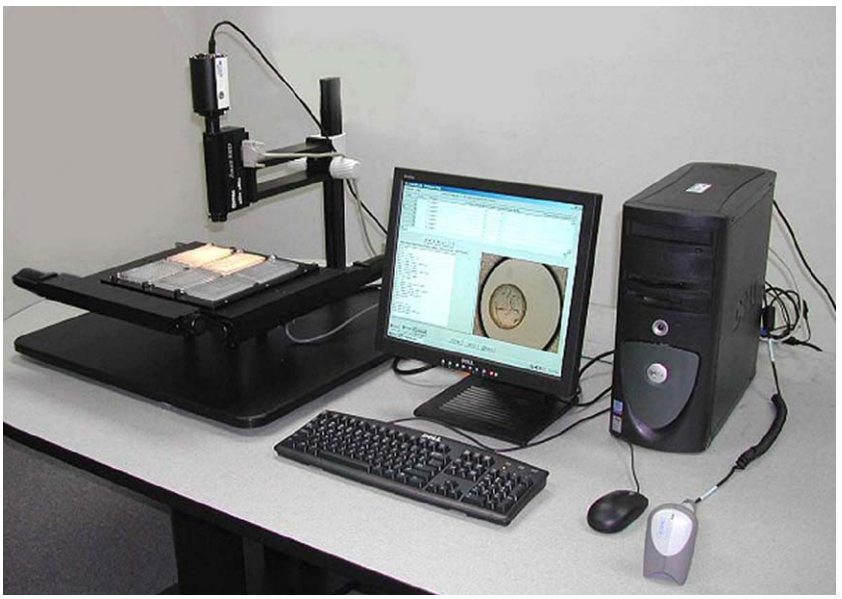

Fig. 2. CrystalScore™.

HTML, Word and Excel, time-lapsed image acquisition, AVI Focus-Through Movies and tray support for Linbro 23, Corning 24 and 96, 48-well, 96-well, Greiner 288-well, Cryschem 24, VDX 24, and Nunc 72-well.

3.4. Predicting protein crystallization conditions using neural network technology

Neural network technology developed from artificial intelligence research was applied to protein crystallization screening and resulted in the ability to accurately predict/recognize conditions that favor crystallization. Preliminary research (developed in collaboration with scientist and engineers from Diversified Scientific Inc., the University of Alabama at Birmingham and Interactive Analysis Inc.) deals with optimization techniques that may increase the success rate for producing diffraction-quality macromolecular crystals (DeLucas et al., 2003). This technology demonstrates the most promise for optimizing protein crystallization if combined with a thorough sampling of "crystallization space". An initial screen based on sampling techniques, such as the incomplete factorial (Carter and Carter, 1979) described previously, is used for the protein crystallization trials. Every crystallization trial outcome, including failures, is used to train a neural network. Once trained, the neural network may recognize conditions that yield crystals. Neural networks are based on a real nervous system paradigm composed of multiple neurons communicating through axon connections. Characteristics of neural networks include self-organization, non-linear processing, and massive parallelism. The neural network exhibits enhanced approximation, noise immunity, and classification properties. The self-organizing and predictive nature of the neural networks allow for accurate prediction of never before seen crystallization conditions, even in the presence of noise. The predictive neural network is trained via back-propagation using the incomplete factorial screen. If properly trained, the neural network can be used to identify or recognize important patterns of crystallization. An input pattern comprised of the incomplete factorial screen is presented to the network. The outputs are compared to the known scores. Additional neurons are added and interconnect weights (basis functions) are adjusted to minimize the error and maximize R^2 between the actual versus the predicted values. This process is continued until the average error across all the training sets is minimized. Eventually, if the correct variables and sample size are chosen to adequately represent the crystallization nature of the protein, a stable set of hidden neurons and basis function weights evolve. This neural network can then be used to predict non-sampled complete factorial conditions to be used for optimization, i.e. predicting the conditions that produce crystals from the entire "crystallization space" of possible experimental conditions based on the results from a much smaller number of actual experiments performed. This approach has a higher probability of producing accurate predictions if the small test set is statistically representative of the "crystallization space".

3.4.1. Previous experiments

Previously, this laboratory reported the successful use of a modified neural network. The results from one protein that was previously screened, 9C9 (*C. elegans* protein expressed as part of the SECSG high-throughput project), will be summarized here. For these initial experiments, the neural network was trained using experiments 1-315 from the complete set of 360 screen conditions. This partial sampling of the incomplete factorial design experiment was used to train a

neural network to recognize conditions that result in crystallization. The neural network trained with all results, including failures. The 315 experiments (used for training) allowed the neural network to converge with an acceptable R^2 value of 0.604. The scoring system was modified from a linear scale with clear drops equal to 0 and crystals scored at 10, to a binary scheme. In the binary scheme any crystalline result was given a mark of 2000, the other results (i.e. clear drop, phase separation, precipitate, microcrystals/precipitate, and rosettes/spherulites) were scored 1–5 respectively. The input to the neural network is the indexed variables and the output is the predicted score. The weights of the hidden neurons are determined by back propagation. The remaining 12.5% (45 experiments) of the incomplete factorial screen results were used for verification. There was only one crystal producing condition in the training set (experiment 239), (Fig. 3). The scoring system used was binary with: non-crystal marked 0 and crystal scored as 2000. Experiments 316–360 were used to verify the neural network. The score or result from a protein crystallization experiment (y-axis) versus the crystallization experiment (x-axis) is displayed. The results from a "real" experiment are plotted in blue and those from the predicted experiment using a neural network are displayed in red. Fig. 4 illustrates the physical outcome of the one predicted crystallization conditions (experiment 350). The trained network was able to predict every crystallization outcome in the 12.5% test set even though the data had never been input to the network. This result in the face of the low R^2 value highlighted the neural network's ability to recognize crystallization conditions for 9C9. The neural network was able to accurately predict the outcomes of the remaining 45 experiments, even though results from these experiments had never been input into the neural network program. An interesting observation that facilitated the 9C9 neural network predictions was that the scoring system emphasized crystals and de-emphasized non-crystals. This analysis, among others not mentioned in this paper, reinforced the hypothesis that the neural network can be used to predict crystallization conditions for previously non-crystallized proteins (DeLucas et al., 2003).

The simplicity of training neural networks and the apparent accuracy of predicting crystallization conditions in such preliminary work had very exciting implications for optimization. It became clear from some of the initial experiments that the ability of the neural network to identify patterns of crystallization in complex non-linear data sets may provide a powerful method of optimization. The total number of permutations possible for a particular

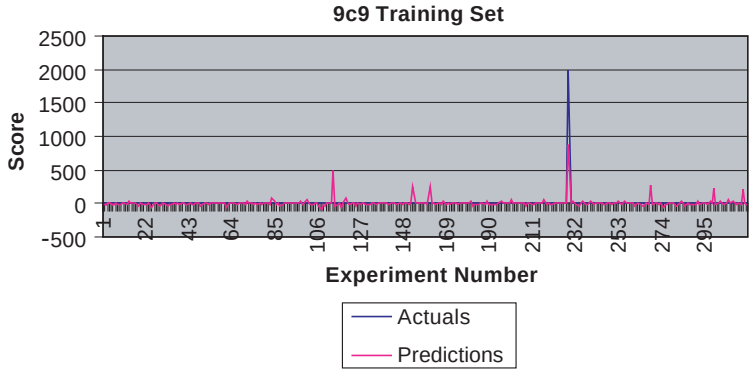

Fig. 3. Neutral network training data for protein 9c9. Experiments 1–315 were used to train the neural network.

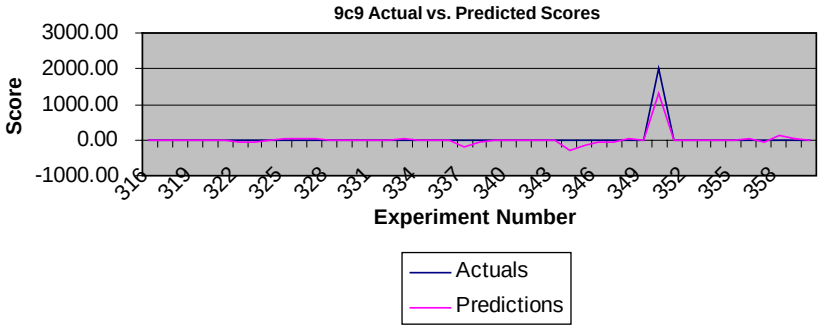

Fig. 4. Comparison of crystallization scores between predicted and actual experiments using input data the neural network has NEVER seen for protein 9c9. Experiments 316–360 were used to verify the neural network that was trained only on experiments 1–315.

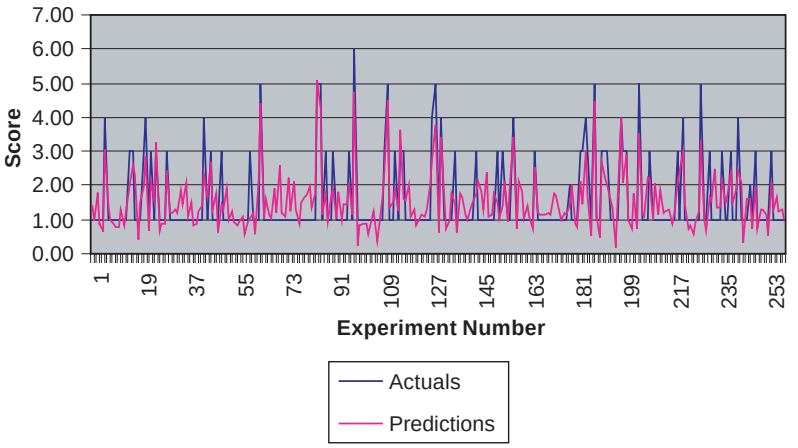

Fig. 5. Typical correlation observed upon completion of neural network training.

screen is calculated by multiplying the number of discrete values of each design variable. For the *C. elegans* protein, 9c9, there are 320,050 possible permutations in the incomplete factorial space.

If actual scores are used for training instead of heavily weighting specific crystal outcomes, there is considerable noise in the predictions. However, the neural network was able to consistently predict the highest scores, in spite of the inclusion of false positives. Fig. 5 shows a typical training pattern observed when all scores are given equal priority to initiate the training. For all recent work described in the following section, scores were assigned equal priority at the outset of training. Although there are false positives, the program does accurately predict the experimental conditions with the highest scores, and the highest predicted values closely match the highest experimental values.

3.4.2. Recent experiments

More recently the original technique of training the network was extended with a modification in the procedure that involved randomly dividing the database into 10 disjointed and unique sets

such that each crystallization condition was used to train the network nine times and for an independent evaluation of the model during a 10th iteration. Each of the 10 sets contained 90% of the data used for training and the remaining 10% was completely withheld from the modeling process and evaluated only after all training was completed. This process gave 10 separate algorithms that captured the variation in the data. The final model was used as a consensus of all the predicted values taken as an average. The standard deviation (sigma) was also calculated as an estimate of the probability of accuracy for an individual prediction. Once the 10 models were derived from the data, all possible combinations of the inputs (~320,000) were calculated, removing the highest 1000 maximum predicted values and sorting these predictions based on the minimum standard deviation. The first 360 conditions (the entire incomplete factorial screen) was used for training with the "trained" neural network and then used to predict non-sampled complete factorial conditions for optimization, i.e. predicting the conditions that produce crystals from the entire "crystallization space". The CBSE experimentally prepared the chemical conditions for the top 360 scores. The number 360 was chosen due to the fact that the Corning crystallization plate holds 120 conditions and the net predicted scores for three different temperatures. Thus, one complete plate of predicted conditions was prepared at each temperature. These selections represented the neural network's choice for the top 360 conditions from 320,000 possibilities for each protein. Interactive Analysis performed all consensus model neural network calculations using back-propagation as described above.

For the recent experiments, 11 proteins were subjected to the incomplete factorial screen followed by neural network analysis as described above. The proteins used included *C. elegans*—9C9, *C. elegans*—11059, variable chitin binding protein-3 (VCBP-3), beta-lactoglobulin (bovine milk), alpha-chymotrypsinogen (bovine pancreas), catalase (bovine liver), collagen binding protein (ACE-40), bacterial hyaluronidase, TSP-1, PA1, and PX1 (proprietary commercial proteins).

The neural network predictive scores for each protein generally ranged between one to three units higher than any of the input scores used for training. Thus, the neural network was able to use the training data to weight those factors determined to be important for optimum crystallization results and subsequently predict new conditions that should produce improved results. Neural network predictive capabilities were compared with linear regression as an alternative method to predict crystallization outcomes. In every case, the neural network was superior to the linear regression algorithm. Fig. 6 demonstrates the typical difference seen for the predictive ability of the neural network versus linear regression analysis.

In some cases the neural network predictions appeared to fall into an area that would be described as part of the region around the initial "crystallization hit". Thus, if a fine screen around the initial "hit" were prepared (without help from the neural net predictions) the improved conditions would also have been found. However, for other proteins (*C. elegans*—11059, VCBP-3, and ACE-40) this clearly was not the case. In these cases, predictions involved an area of "crystallization space" that was significantly different from the initial screening (training) hits and that would not have been explored in a second round of testing. For example, it is highly unlikely that a trained crystallographer would have optimized the initial "hit" for *C. elegans*—11059 at the conditions predicted by the neural network (the pH's differed by 3.0 units and completely different salts and concentrations were used). Fig. 7 and Table 3 show crystals and chemical conditions obtained from the initial screen (best "hit" from initial screen) versus the best

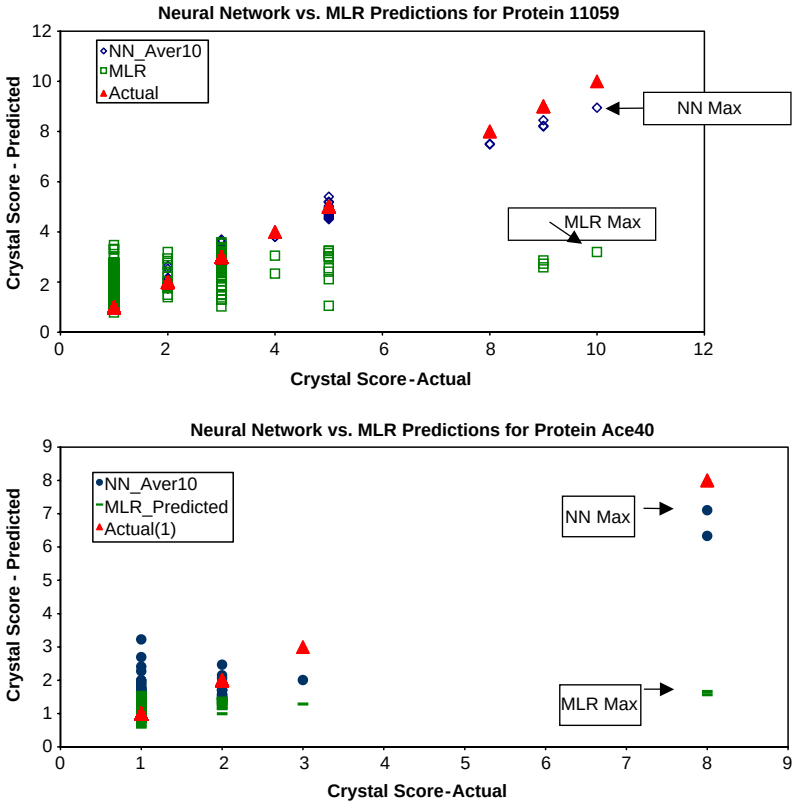

Fig. 6. Scatter plots of multivariant linear regression (—) MLR and neural net (●) (NN) predictions versus actual (▲).

experimental results obtained via the neural network predictions. In some instances the neural network predictions yielded experimental results scored as high as a 10 (one large, high-quality crystal) using the Hampton scoring system. There are two traditional methods used to optimize crystallization based on the results of an incomplete factorial screen. One is to proceed directly to a fine screen centered close to the conditions that produce crystals, maintaining a narrow range of pH, and concentration variables. Usually the identities of the PEG's and anionic salt components would remain constant. A second approach would be to use a linear prediction method to analyze the overall results to determine the best value for each variable. For example, if the average score of experiments containing malonate is higher than the average score for the other salts, malonate would be selected as the anionic salt. The neural network analysis predicts high scores for many experiments in the areas of the original screen "hits". In our set of test proteins, it also included most of the experiments suggested by the linear analysis. Larger and better quality crystals were found for the majority of the proteins screened among the experiments suggested by the neural network analysis rather than those in the original screen (as well as those suggested by the linear analysis). In almost all cases, the linear regression predictive analysis experiment either failed to predict conditions that led to crystals or the crystals were of poorer quality than the original "hits".

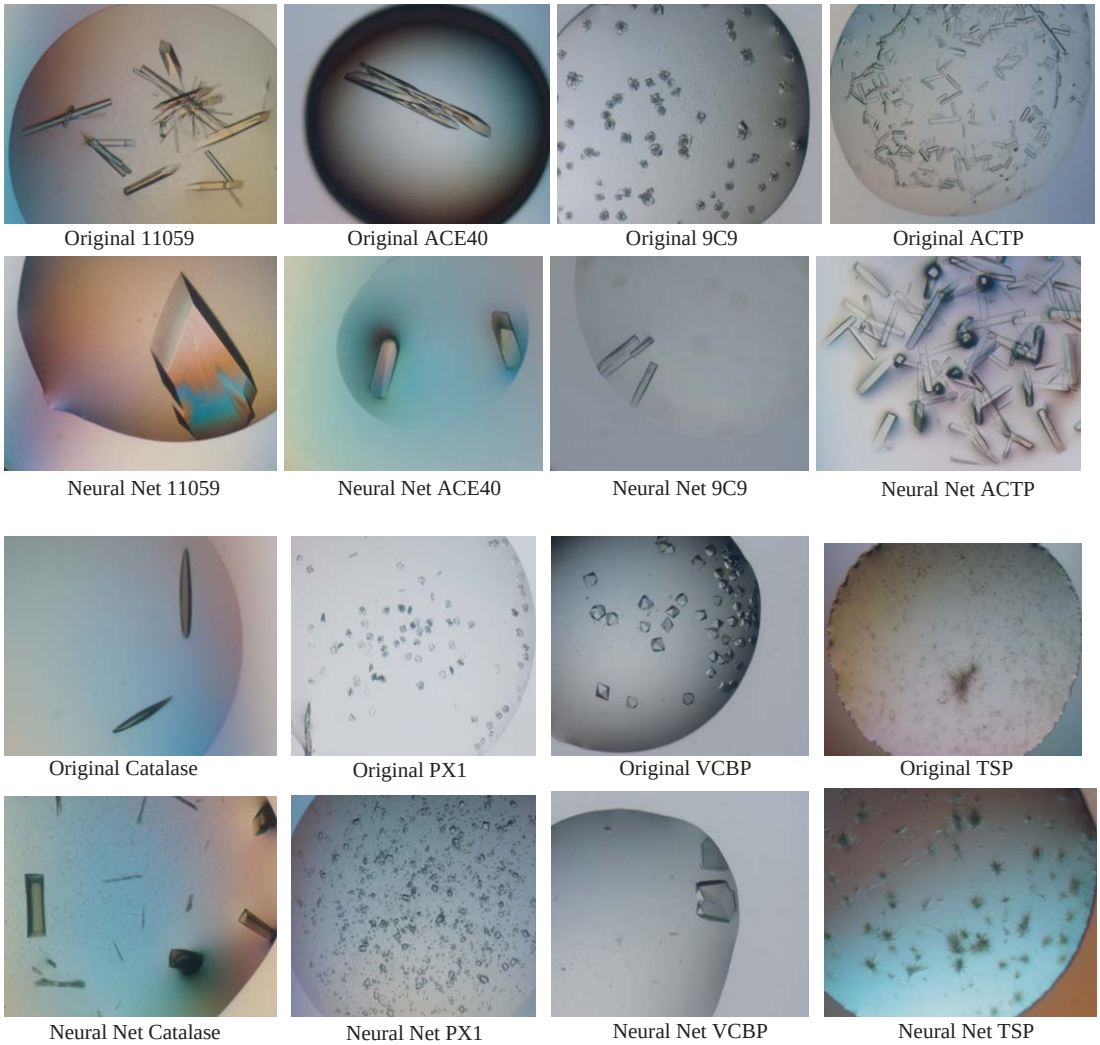

Fig. 7. Crystals images. The crystalline images show the best crystals that were grown with the incomplete factorial screen compared with the best crystals obtained from conditions predicted by the neural network.

These results imply that the neural network may be useful for optimizing the fine screening that is typically required to advance from initial "hits" to diffraction-quality crystals. But it may also be useful for finding new permutations of the components and their concentrations that also yield superior crystals, yet lie far outside the area that would normally be subjected to a fine screen (this area may produce crystals with new space groups or morphologies). Comparison of the actual data versus neural network predictions clearly indicates that the neural network can be improved further by adding other variables (i.e. second virial coefficient, isoelectric point, polydispersity levels, etc.).

Table 3
Corresponding chemical conditions for crystals shown in Fig. 7

Original 11059 0.1 M Acetate, pH 4.5 0.358 M Na acetate 6.8% PEG 400 6% Glycerol 0.01 M $MgCl_2$ 0.05% BOG	Original ACE40 0.1 M Acetate, pH 4.5 0.188 M Na chloride 15% PEGM5000 0.05 M ArgHCl 0.01 M $CaCl_2$	Original 9C9 0.1 M Bicine, pH 8.3 0.289 M Na chloride 23.2% PEGM5000 0% Glycerol 0.01 M $MgCl_2$ 0.05% BOG	Original ACTP 0.1 M Acetate, pH 4.5 0.599 M NH_4 sulfate 11% PEG 1450 6% Glycerol 0.05 M ArgHCl
Neural network 11059 0.1 M HEPES, pH 7.5 0.907 M Na malonate 1.3% PEG 4000 0% Glycerol 0.01 M $MgCl_2$ 0.05% BOG	Neural network ACE40 0.1 M Acetate, pH 4.5 0.716 M Na chloride 1.3% PEG 8000 0% Glycerol 0.01 M $MgCl_2$	Neural network 9C9 0.1 M HEPES, pH 7 0.21 M NH_4 citrate 16% PEGM5000 0% Glycerol 0.01 M $MgCl_2$ 0.05% BOG	Neural network ACTP 0.1 M MES, pH 6 0.21 M Na acetate 17.2% PEG 4000 3% Glycerol 0.01 M $CaCl_2$
Original catalase 0.1 M Bicine, pH 9 0.264 M Na malonate 14.7% PEGM5000 0% Glycerol 0.01 M $MgCl_2$	Original PX1 0.1 M Bicine, pH 8.3 0.264 M Na citrate 4.9% PEG 1450 3% Glycerol 0.01 M $MgCl_2$ 0.05% BOG	Original VCBP 0.1 M HEPES, pH 7.5 0.20 M Na acetate 16% PEG 8000 6% Glycerol 0.05% BOG	Original TSP-1 0.1 M Bicine, pH 9 0.20 M NH_4 sulfate 16% PEGM5000 0.05 M ArgHCl 0.01 M $MgCl_2$
Neural network catalase 0.1 M HEPES, pH 7.0 0.20 M Na malonate 16% PEGM5000 0% Glycerol 0.01 M $CaCl_2$ 0.05% BOG	Neural network PX1 0.1 M HEPES, pH 7.5 0.11 M Na chloride 8.9% PEGM5000 3% Glycerol 0.05% BOG	Neural network VCBP 0.1 M Bicine, pH 9 0.758 M Na acetate 1.1% PEG 1450 6% Glycerol 0.01 M $CaCl_2$ 0.05% BOG	Neural network TSP-1 0.01 M Bicine, pH 8.3 0.239 KSCN 19.2% PEGM5000 3% Glycerol 0.01 M $MgCl_2$ 0.05 M ArgHCl

Table 4 shows the actual scores for the top 120 conditions predicted by the neural network for each of the three temperatures used in the screen. It should be noted that the neural network prediction for each condition ranged between 5 and 10, yet the actual scores fall in a much broader range (1–10). Possible explanation for the disparity between predicted versus actual scores include false positive predictions, the general variability of crystallization (delayed nucleation, contaminant particles in droplets, etc.), protein degradation during the time period between the initial screen, and the neural network optimization. Section 4 describes some of the approaches being used to enhance the neural net's predictive capability for crystallization.

4. Conclusions and future enhancements

The use of a statistically representative crystallization screen may provide an advantage over commercially available screens, particularly for those proteins that crystallize under experimental

Table 4
Summary of neural network predictions. This table provides the total number of conditions predicted for each score (scores 1–10) by the neural network for each of three temperatures: 4, 15, and 22°C.

Temperatures															
	1			2			3			4			5		
	4 °C	15 °C	22 °C	4 °C	15 °C	22 °C	4 °C	15 °C	22 °C	4 °C	15 °C	22 °C	4 °C	15 °C	22 °C
Proteins															
9C9	5	23	10	7	1	0	8	6	11	0	0	0	64	60	83
11O59	57	50	60	23	4	0	16	36	43	0	0	0	3	12	6
PX1	15	10	13	9	0	0	77	78	67	5	9	5	1	2	7
B-lacto	56	45	57	55	67	51	8	4	9	0	0	0	0	0	0
Ace 40	100	106	107	7	3	3	4	4	4	0	0	0	4	4	13
Hya	120	116	117	0	0	0	0	4	3	0	0	0	0	0	0
ACTP	43	39	41	0	0	0	3	2	1	4	0	0	0	0	0
VCBP	94		64	0		0	4		23	0		5	6		0
PA1	120	97	90	0	1	1	0	22	30	0	0	0	0	0	0
Catalase	5	3	4	0	0	0	9	7	16	4	0	1	0	0	0
TSP-1	119	104	106	0	1	1	1	8	4	0	1	0	0	3	2
	6			7			8			9			10		
	4 °C	15 °C	22 °C	4 °C	15 °C	22 °C	4 °C	15 °C	22 °C	4 °C	15 °C	22 °C	4 °C	15 °C	22 °C
Proteins															
9C9							11	8	5	3	2	0	1	0	0
11O59	10	8	2	11	11	8	12	0	0	2	11	2	0	2	3
PX1	0	0	0	6	2	5	7	14	12	0	0	1	0	0	0
B-lacto	4	1	5	1	5	10	0	0	0	0	0	0	0	0	0
Ace 40	0	0	0	0	0	0	3	0	0	2	0	0	0	0	0
Hya	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ACTP	0	0	0	0	0	0	41	21	40	25	44	21	0	0	0
VCBP	5	9	8	0	0	3	7		17	0		7	0		0
PA1	0		0	7		1	0	0	0	0	0	0	0	0	0
Catalase	0	0	0	0	0	0	30	54	50	30	27	31	0	0	0
TSP-1	18	5	1	22	21	16	0	0	0	0	0	0	0	0	0
	0	3	4	0	0	0									

conditions that are uncommon. In addition, the use of such a screen with predictive algorithms may provide a powerful tool for crystallizing and optimizing new proteins. The incorporation of these tools with a nanoliter dispensing system enables the experimentalist to efficiently and intelligently search for the optimum crystallization conditions.

Several improvements to the *NanoScreen*™ system and experimental screens will be implemented in the future. These include increasing throughput by adding tips and dispensing heads to the *NanoScreen*™ System, and improving the effectiveness of our statistical screens. The implementation of other embodiments of original reduced-volume screening concepts to further

improve the efficiency and effectiveness of screening protein crystallization conditions are also being pursued. These include customizing tip dispense rates to accommodate more viscous PEG solutions and more flexible control software that automatically calculates and prepares neural network optimization experiments.

We anticipate the predictive output of a neural network could be significantly improved if an information rich phenomenological parameter, such as the second virial coefficient, was used as input in addition to the empirical parameters. The second virial coefficient, B, is a measure of two body (protein–protein) interactions in a defined solution condition. The importance of B as a predictor with regard to protein crystallization has been well established both experimentally and theoretically (George and Wilson, 1994; George et al., 1997; Ducruix et al., 1996; Malfois et al., 1996; Rosenbaum et al., 1996; Neal et al., 1998; Neal et al., 1999; Bonnete et al., 1999). A "crystallization slot" is correlated with solution conditions that are favorable to crystallization and corresponds to B values in the range of about -1×10^{-4} to -8×10^{-4} (mol ml g^{-2}) and indicates protein–protein interactions that are slightly to moderately attractive. Protein crystal growth (PCG) trials conducted in solution conditions at more negative B values have a greater risk of forming amorphous solid phase because of corresponding stronger protein–protein attractions. On the other hand, experiments at more positive values, where the network protein–protein interactions are repulsive, typically require protein concentrations that are impractically high to cause phase separation of any kind. The crystallization slot can be used as an effective guide by crystallographers to direct changes in a particular solution parameter (pH, temperature, crystallization agent and concentration, etc.) that will increase the probability of a successful crystallization trial. It is important to point out that conducting PCG experiments under conditions that correspond to the crystallization slot does not guarantee a successful crystallization trial. However, working at conditions outside the slot most assuredly reduces the probability of a desirable outcome.

An additional feature of B regarding protein crystallization is its ability to mimic solubility, s, behavior of a protein (George et al, 1997; Gripon et al., 1997; Guo et al., 1999; Demoruelle et al., 2002). This finding suggests that the protein–protein pair potentials manifest in undersaturated (even dilute) protein solutions extend into supersaturation regions. Subsequent studies (Rosenbaum and Zukoski, 1996) showed that the lysozyme phase boundary could be constructed using an adhesive hard sphere potential along with B values to model protein interactions. The theoretical basis for the direct link between B and s has been presented (Haas et al., 1999; Ruppert et al., 2001). Major findings of these studies show that protein interactions were strongly anisotropic and that crystallization conditions had little effect on the interaction distance or the anisotropy between the protein molecules. The correlation between B and s offers crystallographers a distinct advantage for determining the solubility behavior of proteins. From the practical standpoint of setting up PCG trials, it is desirable to know how a protein's solubility depends on a particular solution parameter. For example, for any rational PCG approach, it is important to know if and to what extent the protein's solubility has a normal, retrograde, or no temperature dependence. In other cases, knowledge of the dependence of s on solution pH, ionic strength, type, and concentration of crystallization agent, etc. will allow systematic PCG trials to be conducted. Since it is impractical to try and determine absolute values for s by equilibrium crystallization studies as a function of so many solution variables, the measurement of B provides a realistic alternative for obtaining the s behavior.

The incorporation of B values as an input parameter for the neural network is an obvious extension towards the development of a more accurate and robust outcome for the network prediction. A distinct advantage for this parameter is that B values can be determined from dilute solution measurements on the protein, i.e. no actual crystallization trials with often ambiguous scoring are required. In addition, B values reflect a phenomenological aspect of PCG (protein–protein interactions and solubility behavior) that has a sound theoretical and experimental basis. However, to make this approach practical, a way must be found to determine B that is relatively fast (minutes per B determination), requires small amounts of protein (microgram or sub-microgram quantities per B determination), and is compatible with a robotic platform for high-throughput experiments. The direct and most often reported way of determining B values to compare with the crystallization slot is by static light scattering (SLS) (George and Wilson, 1994) in which scattered intensity versus protein concentration data are collected. Although SLS is the traditional method of choice for B measurements, its operational features are not consistent with the above-listed requirements for high-throughput experiments. A primary disadvantage of SLS is that the amount of protein sample needed for a single B measurement is at a minimum typically on the order of hundreds of micrograms, depending on the protein size. To circumvent the disadvantages of SLS for B measurements using a high-throughput platform, an alternative approach utilizing microchip self-interaction chromatography (SIC) has been presented (Garcia et al., 2003). The results for this work showed that the chromatographic retention times by SIC were highly correlated with B measurements by SLS. A single B measurement by SIC was performed in less than 10 min with as small a protein sample as 37 ng, and the chromatographic platform makes SIC compatible with high-throughput configurations already commercially available. Thus, the utilization of a microchip SIC approach can potentially make the task of B measurements more routine so that providing B values as input to the neural network will become a standard protocol.

The success of a neural network's ability to predict protein crystallization ultimately depends on the validity of the input used for the training. First and foremost, the variables used to design the crystallization recipes must encompass the target protein's crystallization domain. If this condition is not met, screening will not result in crystalline outputs. Second, the tremendous effort of artificial intelligent research has resulted in the development of numerous types of neural networks that are superior in predicting outcomes compared to networks trained using simple back-propagation. Many of these neural networks use a probalistic approach to weighting individual nodes. These networks form conditional networks or ''belief'' networks. The individual results from a crystallization screen can be considered events and the probability of crystallizing a protein calculated as the ratio of number of desired outcomes per total number of outcomes. In a very real sense there is always a level of uncertainty associated with sampling a protein's ''crystallization space''. The level of uncertainty in screening protein crystallization space can be modeled using a probabilistic inference network. Empirical data, such as results from an incomplete factorial screen described above, can be used as an input pattern in conjunction with other data, such as the isoelectric point of a protein, molecular weight, helical content, second virial coefficient, protein source, etc. The additional empirical data representing calculated and measured biophysical properties plus the known crystallization behavior of the protein can be used to improve the training by modeling the uncertainties associated with a particular target. It is anticipated that a database of such information could be

mined to discover correlations between biophysical properties of a molecule and its crystallization behavior.

Acknowledgements

Funding for this research was provided by the NIH Protein Structure Initiative Grant P50-GM62407 and the NASA Cooperative Agreement NCC 8-246.

References

Ay, J., Gotz, F., Borriss, R., Heinemann, U., 1998. structure and function of the *Bacillus* hybrid enzyme GluXyn-1: native-like jellyroll fold preserved after insertion of autonomous globular domain. Proc. Natl. Acad. Sci. USA 95, 6613–6618.

Baker, H.M., Day, C.L., Norris, G.E., Baker, E.N., 1994. Enzymatic proteins. Acta Crystallogr. D 50, 380–384.

Betton, J.M., Jacob, J.P., Hofnung, M., Broome-Smith, J.K., 1997. Creating a bifunctional protein by insertion of beta-lactamase into the maltodextrin-binding protein. Nat. Biotechnol. 15, 1276–1279.

Bonnete, F., Finet, S., Tardieu, A., 1999. Second virial coefficient: variations with lysozyme crystallization conditions. J. Cryst. Growth 196, 403–413.

Campbell, J.W., Duee, E., Hodgson, G., Mercer, W.D., Stammers, D.K., Wendell, P.L., Muirhead, H., Watson, H.C., 1972. X-ray diffraction studies on enzymes in the glocolytic pathway. Cold Spring Harbor Symp. Quant. Biol. 36, 165–170.

Carter Jr., C.W., Carter, C.W., 1979. Protein crystallization using incomplete factorial experiments. J. Biol. Chem. 254, 12219–12223.

Charron, C., Kern, D., Giege, R., 2002. Crystal contacts engineering of aspartyl-tRNA synthetase from *Thermus thermophilus*: effects on crystallizability. Acta Crystallogr. D 58, 1729–1733.

Chayen, N.E., 1997. The role of oil in macromolecular crystallization. Structure 5 (10), 1269–1274.

Chen, P., Tsuge, H., Almassy, R.J., Gribskov, C.L., Katoh, S., Vanderpool, D.L., Margosiak, S.A., Pinko, C., Matthews, D.A., Kan, C.C., 1996. Structure of the human cytomegalovirus protease catalytic domain reveals a novel serine protease fold and catalytic triad. Cell 86, 835–843.

D'Arcy, A., Stihle, M., Kostrewa, D.A., Dale, G., 1999. Crystal engineering: a case study using the 24 kDa fragment of the DNA gyrase B subunit from *Escherichia coli*. Acta Crystallogr. D 55, 1623–1625.

Dale, G.E., Broger, C., Langen, H., D'Arcy, A., Stuber, D., 1994. Improving protein solubility through rationally designed amino acid replacements: solubilization of the trimethoprim-resistant type S1 dihydrofolate reductase. Protein Eng 7, 933–939.

Davis, S.J., Brady, R.L., Barclay, A.N., Harlos, K., Dodson, G.G., Williams, A.F., 1990. Crystallization of a soluble form of the rat T-cell surface glycoprotein CD4 complexed with Fab from the W3/25 monoclonal antibody. J. Mol. Biol. 213, 7–10.

DeLucas, L.J., Bray, T.L., Nagy, L., McCombs, D., Chernov, N., Hamrick, D., Cosenza, L., Belgovskiy, A., Stoops, B., Chait, A., 2003. Efficient protein crystallization. J. Struct. Biol. 142, 188–206.

Demoruelle, K., Guo, B., Kao, S., McDonald, H., Nikic, D., Holman, S., Wilson, W., 2002. Correlation between the osmotic second virial coefficient and solubility for equine serum albumin and ovalbumin. Acta Crystallogr. D 58, 1544–1548.

DeTitta, G.T., Biance, M.A., Collins, R.J., Faust, A.M.E., Kaczmarek, J.N., Luft, J.R., Fehrman, N.A., Pangborn, W.A., Salerno, J.M., Wolfley, J.R., 2001. Macromolecular crystallization in a high throughput setting. Conference and Exhibit on International Space Station Utilization, October 15–18, Cape Canaveral, FL.

Ducruix, A., Guilloteau, J., Ries-Kautt, M., Tardieu, A., 1996. Protein interactions as seen by solution X-ray scattering prior to crystallogenesis. J. Cryst. Growth 168, 28–39.

Garcia, C., DeGail, H., Wilson, W., Henry, G., 2003. Measuring protein interactions by microchip self-interaction chromatography. Biotech. Progress 19, 1006–1010.

George, A., Wilson, W., 1994. Predicting protein crystallization from a dilute solution property. Acta Crystallogr. D 50, 361–365.

George, A., Chiang, Y., Guo, B., Arabshahi, A., Cai, Z., Wilson, W., 1997. Second virial coefficient as a predictor in protein crystal growth. Methods Enzymol 276, 100–110.

Goulding, C.W., Perry, L.J., 2003. Protein production in *E. coli* for structural studies by X-ray crystallography. J. Struct. Biol. 142, 133–143.

Gripon, C., Legrand, L., Rosenman, I., Vidal, O., Robert, M., Boue, F., 1997. Lysozyme–lysozyme interactions in under-and supersaturated solutions: a simple relation between the second virial coefficients in H_2O and D_2O. J. Cryst. Growth 178, 575–584.

Grueninger-Leitch, F., D'Arcy, A., D'Arcy, B., Chene, C., 1996. Deglycosylation of proteins for crystallization using recombinant fusion protein glycosidases. Protein Sci 12, 2617–2622.

Guo, B., Kao, S., McDonald, H., Asanov, A., Combs, L., Wilson, W., 1999. Correlation of second virial coefficients and solubilities useful in protein crystal growth. J. Cryst. Growth 196, 424–433.

Haas, C., Drenth, J., Wilson, W., 1999. Relation between the solubility of proteins in aqueous solutions and the second virial coefficient of the solution. J. Phys. Chem. B 103, 2808–2811.

Hofmeister, F., 1888. On the understanding of the effects of salts, Arch. Exp. Pathol. Pharmakol. (Leipzig) 24, 247–260.

Hosfield, D., Palan, J., Hilgers, M., Scheibe, D., McRee, D.E., Stevens, R.C., 2003. A fully integrated protein crystallization platform for small-molecule drug discovery. J. Struct. Biol. 142, 207–217.

Jancarik, J., Kim, S.-H., 1991. Sparse matrix sampling: a screening method for the crystallization of macromolecules. J. Appl. Crystallogr. 24, 409–411.

Kostrewa, D., Grueninger-Leitch, F., D'Arcy, A., Broger, C., Mitchell, D., van Loon, A.P.G.M., 1997. Crystal structure of phytase from *Aspergillus ficuum* at 2.5 Å resolution. Nat. Struct. Biol. 4, 185–190.

Krupka, H.I., Rupp, B., Segelke, B.W., Lekin, T.P., Wright, D., Wu, H.-C., Todd, P., Azarani, A., 2002. The high-speed Hydra-Plus-One system for automated high-throughput protein crystallography. Acta Crystallogr. D 58, 1523–1526.

Lawson, D.M., Artymiuk, P.J., Yewdall, S.J., Smith, J.M., Livingstone, J.C., Treffry, A., Luzzago, A., Levi, S., Arosio, P., Cesareni, G., 1991. Solving the structure of human H ferritin by genetically engineering intermolecular crystal contacts. Nature 349, 541–544.

Longenecker, K.L., Garrard, S.M., Sheffield, P.J., Derewenda, Z.S., 2001. Protein crystallization by rational mutagenesis of surface residues: Lys to Ala mutations promote crystallization of RhoDGI. Acta Crystallogr. D 57, 679–688.

Luft, J.R., Wolfley, J., Jurisica, I., Glasgow, J., Fortier, S., DeTitta, G.T., 2001. Macromolecular crystallization in a high throughput laboratory—the search phase. J. Cryst. Growth 232, 591–595.

Malfois, M., Bonnete, F., Belloni, L., Tardieu, A., 1996. A model of attractive interactions to account for fluid–fluid phase separation of protein solutions. J. Chem. Phys. 105, 3290–3300.

Mateja, A., Devedjiev, Y., Krowarsch, D., Longenecker, K., Dauter, Z., Otlewski, J., Derewenda, Z.S., 2002. The impact of GluAla and GluAsp mutations on the crystallization properties of RhoDGI: the structure of RhoDGI at 1.3 Å resolution. Acta Crystallogr. D 58, 1983–1991.

McElroy, H.E., Sisson, G.W., Schoettlin, W.E., Aust, R.M., Vallafranca, J.E., 1992. Studies on engineering crystallizability by mutation of surface residues of human thymidylate synthase. J. Cryst. Growth 122, 265–272.

McPherson, A., 1982. Preparation and Analysis of Protein Crystals. Wiley, New York.

Mueller, U., Nyarsik, L., Horn, M., Rauth, H., Przewieslik, T., Saenger, W., Lehrach, H., Eickhoff, H., 2001. Development of a technology for automation and miniaturization of protein crystallization. J. Biotechnol. 85, 7–14.

Nagi, A.D., Regan, L., 1997. An inverse correlation between loop length and stability in a four-helix-bundle protein. Fold Des 2, 67–75.

Neal, B., Asthagiri, D., Lenhoff, A., 1998. Molecular origins of osmotic second virial coefficients of proteins. Biophys. J. 75, 2469–2477.

Neal, B., Asthagiri, D., Velov, O., Lenhoff, A., Kaler, E., 1999. Why is the osmotic second virial coefficient related to protein crystallization? J. Cryst. Growth 196, 377–387.

Nugent, P.G., Albert, A., Orprayoon, P., Wilsher, J., Pitts, J.E., Blundell, T.L., Dhanaraj, V., 1996. Protein engineering loops in aspartic proteinases: site-directed mutagenesis, biochemical characterization and X-ray analysis of chymosin with a replaced loop for rhizopuspepsin. Protein Eng 9, 884–893.

Oefner, C., D'Arcy, A., Hennig, M., Winkler, F.K., Dale, G.E., 2000. Structure of human neutral endopeptidase (Neprilysis) complexed with phosphoramidon. J. Mol. Biol. 296, 341–349.

Ostermeier, C., Harrenga, A., Ermler, U., Michel, H., 1997. Structure at 2.7 Å resolution of the *Paracoccus denitrificans* two-subunit sytochrome c oxidase complexed with an antibody FV fragment. Proc. Natl. Acad. Sci. USA 94, 10547–10553.

Prongay, A.J., Smith, T.J., Rossmann, M.G., Ehrlich, L.S., Carter, C.A., McClure, J., 1990. Preparation and crystallization of a human immunodeficiency virus p24–Fab complex. Protein Eng 7, 933–939.

Rosenbaum, D., Zukoski, C., 1996. Protein interactions and crystallization. J. Cryst. Growth 169, 752–758.

Ruppert, S., Sandler, S., Lenhoff, A., 2001. Correlation between the osmotic second virial coefficient and the solubility of proteins. Biotech. Progress 17, 182–187.

Rosenbaum, D., Zamora, P., Zukoski, C., 1996. Phase behavior of small attractive colloidal particles. Phys. Rev. Lett. 1, 150–153.

Santarsiero, B.D., Yegian, D.T., Lee, C.C., Spraggon, G., Gu, J., Scheibe, D., Uber, D.C., Cornell, E.W., Nordmeyer, R.A., Kolbe, W.F., Jin, J., Jones, A.L., Jaklevic, J.M., Schultz, P.G., Stevens, R.C., 2002. An approach to rapid protein crystallization using nanodroplets. J. Appl. Crystallogr. 35, 278–281.

Stevens, R.C., 2000. High throughput protein crystallization. Curr. Opin. Struct. Biol. 10, 558–563.

Thompson, M.J., Eisenberg, D., 1999. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. J. Mol. Biol. 290, 595–604.

Waldo, G.S., Standish, B.M., Berendzen, J., Terwilliger, T.C., 1999. Rapid protein-folding assay using green fluorescent protein. Nat. Biotechnol. 17, 691–695.

Zhou, H.X., Hoess, R.H., DeGrado, W.F., 1996. In vitro evolution of thermodynamically stable turns. Nat. Struct. Biol. 3, 446–451.