Inhibition of protein crystallization by evolutionary negative design

Jonathan P. K. Doye,* Ard A. Louis,† and Michele Vendruscolo‡
University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, United Kingdom
(Dated: February 9, 2008)

In this perspective we address the question: why are proteins seemingly so hard to crystallize? We suggest that this is because of evolutionary negative design, i.e. proteins have evolved not to crystallize, because crystallization, as with any type of protein aggregation, compromises the viability of the cell. There is much evidence in the literature that supports this hypothesis, including the effect of mutations on the crystallizability of a protein, the correlations found in the properties of crystal contacts in bioinformatics databases and the positive use of protein crystallization by bacteria and viruses.

PACS numbers: 87.15Nn,87.14Ee

The overwhelming impression one gets from reading the literature on protein crystallization and listening to experts is that protein crystallization is difficult and requires considerable effort. Furthermore, experience and a certain feeling for what might work can play a crucial role. Recent technical innovations,[1] such as the availability of scanning kits which codify experience to scour for appropriate crystallization conditions, have helped to provide valuable savings in labour. These advances, however, have not altered what seems to be the basic fact: Proteins, for the most part, do not seem to want to crystallize, and have to be coaxed into doing so through the use of suitable cunning.

This situation is particularly vexing, because protein crystallization is a vital step in protein structure determination, and hence to structural genomics initiatives,[2] which seek to catalogue the protein structures associated with the whole genome of a target organism. Although there are also obstacles associated with the expression and purification of the proteins, crystallization is often labelled as the major bottleneck in this process.[3]

The quantification of some of the difficulties involved in protein crystallization is beginning to emerge from structural genomics pilot studies. Generally, the output of new protein structures so far has been "disappointingly low".[4] For example, for a thermophilic prokaryote, probably the class of organisms for which the greatest success rate is expected, only 13% of a target set of non-membrane proteins were estimated to be readily amenable to structural determination; at present only 4% of the structures of these proteins have actually been obtained.[5] These successes probably represent the "low-hanging fruits" of the proteome. How to reach higher branches remains unclear.

In this perspective, we would like to take a step back and offer our opinions on an important question raised by this situation: Why is the crystallization of proteins so difficult? This is not only a fundamental question, but also a practical one. A natural starting point for any rational attempt to overcome the obstacles that hinder protein crystallization is to first understand the nature of these barriers.

In general, one expects that it should be possible to obtain crystals for soluble molecules that have a well-defined structure.[6] So why should globular proteins be any different? One possible answer is that proteins are polypeptide chains with significant conformational entropy and this will have some effect on their crystallization properties. However, their dynamic nature does not interfere with their ability to form specific complexes with proteins and other molecules.

In our opinion, the answer to this question lies in the evolutionary origin of proteins. Proteins are a very special type of polymer and their possible states are different from those of normal polymers. For example, simple homopolymers can be either in a swollen or a collapsed phase, depending on the quality of the solvent.[7] But whereas proteins in a collapsed globular state can remain soluble for appreciable concentrations, collapsed homopolymers aggregate very easily. There are, of course, many more differences between simple polymers and proteins. Here we suggest that evolution appears to have enhanced the tendency to keep globular proteins soluble and active, reducing the probability of realizing all types of aggregate states.

Our hypothesis is thus that proteins have evolved not to crystallize, because crystallization, as well as any type of aggregation, compromises the viability of the cell. Most aggregation diseases, e.g. Alzheimer's and Creutzfeldt-Jakob disease, are associated with non-native protein structures, and the cell has developed sophisticated quality control mechanisms to cope with misfolded proteins.[8] However, there are also a number of diseases associated with the aggregation of proteins in their native state. Perhaps the best known example is sickle cell anaemia, where a mutant form of hemoglobin coalesces to form ordered fibrillar aggregates inside red blood cells. In addition, there are also instances of diseases that result from crystallization: Certain forms of cataracts and anaemia are caused by crystallization of mutant forms of the γ crystallin[9] and hemoglobin[10] proteins, respectively. Furthermore, protein crystallization has been found to be associated with other pathologies[11]. In general, however, such diseases are less common that those associated with the aggregation of misfolded proteins. We suggest that this difference is because the well-defined structure of the

native state makes it much more amenable to evolutionary control.

One further consideration is that the selection pressure is with respect to crystallization *in vivo*, whereas protein crystallographers explore far-from-physiological conditions *in vitro*. However, in our view, the fact that crystallization is difficult even in the latter circumstances simply reflects the robustness of the strategies used by nature to ensure that proteins do not crystallize in the cellular environment.

Our hypothesis is one example of a negative design principle. More often we think in terms of positive design, i.e. that the sequence of a protein has been optimized through evolution to give the protein particular characteristics. However, negative design leading to the avoidance of unwanted properties, such as crystallizability or aggregation, can be equally important.

Such negative design principles have been previously proposed for both the single-molecule and intermolecular properties of proteins. For example, for a protein to fold reliably to its native state, not only must the native structure be particularly low in free energy, but alternative conformations must also not have similar or lower stability.[12] Some of the strategies by which this specificity can be achieved have been identified and then applied in the *de novo* design of proteins.[13] For example, even though it is generally more thermodynamically favourable to have hydrophobic residues in the core of the protein, greater specificity can be achieved by the introduction of some interacting polar residues into the core.[14]

Lessons on negative design can be learnt from the necessity to avoid aggregation. This is a particular problem for proteins involving β-sheets, since their edges are natural sites for association with other β-sheets in nearby proteins, and, for example, can lead to the extended β-sheet structures found in amyloid deposits. A number of negative design strategies have been found in natural proteins that protect β-sheet edges.[15] The simplest strategy is to form a continuous β-sheet structure without any edges, as in β-barrels. Another of the identified strategies has been successfully applied to turn an aggregating protein into a soluble monomeric form by a single mutation of a non-polar residue to lysine.[16]

Designing out unwanted interactions is also necessary in molecular recognition. To achieve specificity, a protein must not only interact strongly with the target molecule, but also have much less favourable interactions with all other molecules.[17,18]

The two examples discussed above illustrate the combination of positive and negative design that is used to tailor the interprotein interactions. Most generally, this is seen in the remarkable properties of cellular solutions, where crowded, multi-component mixtures with protein packing fractions of up to 40%[19] can be both functionally active and stable. By contrast, any attempt to make artificial nanocolloidal mixtures of similar density is bound to result in components sticking together to form an amorphous deposit. In fact, colloid scientists expend considerable effort modifying the surfaces of colloids—adding, for example, charged groups or short polymer brushes—to prevent this from occurring. To achieve this combination of specific attraction (positive design) and generic repulsion (negative design), evolution must exert remarkable control over the matrix of all possible interprotein interactions.[20,21] In this context, our hypothesis concerns a particular type of interaction (namely crystal-forming) that contributes to the diagonal elements (i.e. self-interactions) of this matrix.

Let us consider how this negative design might be achieved. As many amino acid sequences can give rise to the same final protein fold, there is considerably freedom in how the amino acids, particularly those on the surface of the protein,[22] are chosen. This flexibility could potentially allow the protein surface to be organized such that crystallization is hindered, without affecting either the structure of the protein's fold or its active site.

Importantly, such a scenario has testable consequences. If the surfaces of proteins have been optimized to sufficiently reduce their crystallizability, one would expect that random mutations of the surface amino acids that do not alter the structure of the protein fold or its activity (i.e. only the 'neutral' mutations that are evolutionarily allowed) would be likely to lead to a more crystallizable protein. By contrast, if our hypothesis did not apply and a protein's crystallizability did not influence the choice of surface amino acids, one would expect such mutants to be as likely to hinder as to enhance a protein's crystallizability.

We know of two such systematic studies of the crystallizability of mutagens, the first on human thymidylate synthase[23] and the second on a fragment of the DNA gyrase B subunit from *Escherichia Coli*.[24] In both studies, mutations were found to have a dramatic effect on the crystallization properties of the protein. In agreement with our negative design hypothesis, the mutants generally showed enhanced crystallizability compared to the wild-type, as measured by the number of hits in a crystallization screen. There was also evidence of enhancement in crystal quality. Moreover, some of the mutants crystallized in space groups that were not encountered for the wild-type protein. Although the amount of data is not enough to provide conclusive justification of our negative design argument, it is strongly suggestive. Furthermore, there is a body of more anecdotal evidence consistent with our ideas, namely the growing catalogue of proteins that have been first crystallized as mutants.[25]

By contrast, where there has been positive design of the protein surface, as in the case of specific functional binding interactions between two proteins, one would expect random mutagenesis to lead on average to a reduction in the binding affinity between the proteins. This is indeed the case, and such studies have played an important role in understanding the nature of protein-protein binding through the identification of small sets of residues that are key to the stability of the interface.[26]

Although it seems clear that the surfaces of proteins have been designed to hinder crystallization, there still remains the question of what physical mechanism underlies the reduced crystallizability of the evolutionary selected protein surfaces. One might guess that this behaviour reflects some complex property of the surface, and hence would be hard to identify or rationally control. However, there is experimental evidence that surface lysine residues could play a key role in this negative design strategy.

As one would expect for a charged amino acid, lysine prefers to be at the surface of the protein, where it can interact with the aqueous environment. In fact, lysine has the highest propensity to be at the surface of all the amino acids and is the most common surface residue.[27] Lysine is also unique in presenting the largest amount of solvent accessible surface area that is hydrophobic in character,[28] because of the long hydrophobic tail that links the amine group to the protein backbone. Even more interestingly for our present considerations, systematic studies of interprotein contacts have found lysine to be the most underrepresented amino acid at crystal contacts,[29,30] and even more so at the interfaces between subunits of protein oligomers[29] and between proteins that form functional complexes.[31,32] These negative correlations of course raise questions concerning the purpose of lysine residues: Why are they so abundant on the surface, if they are only reluctantly involved in functional interactions? It could be that lysine plays an important negative role in regulating interprotein interactions through preventing unwanted interactions. Indeed, Dasgupta *et al.* suggested the mutation of lysine residues as a rational strategy for enhancing crystallizability.[29]

Just such an approach has been implemented in the experiments of the Derewenda group.[33,34,35,36] They considered the effects of a series of lysine to alanine mutations for human RhoGDI.[33] Their rationale for this particular type of mutation was that the substitution of an amino acid with high conformational entropy by a smaller one would lead to a reduction in the entropy loss on crystal contact formation. Whether for this reason or not—the replacement of a charged amino acid by a neutral one will also lead to concomitant changes in the electrostatic interactions—the results were dramatic. The mutants invariably showed enhanced crystallizability, and often produced crystals that diffracted to higher resolution than achievable otherwise. Consistent with the idea that the lysine residues somehow prevent unwanted interactions, new crystal contacts were often formed at the sites of the mutations. A similar study on glutamate to alanine mutations also revealed enhanced crystallizability, although not quite to the same degree.[36] This rational mutagenesis strategy has since been successfully applied to crystallize proteins of previously unknown structure.[34,35]

Additional support for the idea that negative design is a key aspect of evolution at the molecular level comes from instances where one of the assumptions of our hypothesis does not hold; namely, that crystallization is harmful to the cell. Although this assumption is likely to be generally true, it is a simplification and will not necessarily hold for all cellular environments. In the absence of such a selection pressure, crystallization is likely to be significantly easier. Indeed, there may even be circumstances when crystallization is a positive advantage. For example, a crystal may provide an efficient and convenient way to store a protein. Anecdotal evidence for this correlation between crystallizability and function can perhaps be found in the history of protein crystallization,[11] as it is reasonable to expect that proteins that were among the first to be crystallized are at the easier end of the spectrum of crystallizability. For example, storage proteins, particularly the globulins found in seeds and nuts, were amongst the earlier protein crystals to be discovered, although this, at least partly, also reflects the ready availability of a protein source.

More direct evidence for this potential positive side to crystallization comes from the identification of crystals *in vivo*, an interesting overview of which is given in Ref. 11. For example, protein crystals have been observed in the egg yolks of various organisms, and ribosome crystals have been found in hibernating animals, presumably because they act as a temporary reservoir for this important cellular component. Particularly interesting in this regard is the *Bacillus thuringiensis* class of bacteria, which produce protein toxins specific to a wide variety of insects.[37] Crystals provide a particularly stable (up to periods of years) form for these bacteria to store these toxins. When ingested, these crystals dissolve, releasing the toxins to attack the gut wall of the target insect, thus facilitating the entry of germinating bacterial spores into the host.

Although perhaps harmful to the host cell, there seems little reason why the formation of crystals of virus particles would be disadvantageous to the virus. Indeed, it probably presents a convenient way to densely pack the particles and so minimize possible constraints on self-replication. Consistent with this supposition, crystals of spherical and icosahedral viruses are frequently observed in infected cells. Furthermore, viruses were also amongst the earlier biological particles to be crystallized.

Even more fascinating is the ingenious use of protein crystallization made by viruses that are able to form a quiescent state by embedding themselves in a protein crystal matrix.[38] These viruses cause large quantities of an easily crystallizable protein to be expressed in an infected cell. Nucleation of crystals of this protein then occurs on the surface of the viral particles, surrounding them by crystal and providing the viruses with a protective environment until further transmission is possible. Similar to the bacterial toxins, these crystals readily dissolve in the gut of the insect host, releasing the virus.

The important lesson from these examples is that when it is beneficial for the organism, nature seems to have no difficulty enabling proteins to crystallize. Indeed, such crystals can form spontaneously in the cell simply when the concentration is sufficiently high without the need

for extremely high purities and a series of precipitants to drive the process. The contrasting difficulty that most proteins have in crystallizing, therefore, does not seem to be an intrinsic property of polypeptide chains that have a well-defined folded structure. Rather, it is a property that has been selected by nature, because of the need for the protein-protein interactions to be strictly controlled if the cell is to function properly.

Our arguments are not undermined by the fact that proteins show a whole spectrum of crystallizabilities, with proteins such as lysozymes, hemoglobins and insulins at the easier end. This is to be expected from our perspective. Firstly, as we have seen, the strength of the selection pressure against crystallization may vary considerably (and even be reversed) depending on the function and environment experienced by the protein. Secondly, evolution has no interest in controlling the properties of proteins in non-physiological conditions, and so one should not expect a uniform response. Instead, the degree to which the *in vivo* low crystallizability carries over to *in vitro* environments is likely to show significant variability. Lastly, evolution just requires the crystallizability to be low enough to pose only a low risk to the cell. But there is no reason why the crystallizability could not be significantly below this threshold value, as long as it is not achieved at the expense of the other properties of the protein.

Because the individual concentrations for the majority of proteins are very low relative to the overall protein concentration, some might argue that the putative negative design acts most directly against the non-specific aggregation of native proteins, and then, perhaps because the mechanisms used are generic, only indirectly against crystallization. Indeed, the evidence that we have presented for negative design with respect to crystallization does not indicate whether this effect is direct or indirect. Moreover, the typical cellular concentration of a protein in the cell will be one of the factors that determines the magnitude of the selection pressure against crystallization. However, it should also be remembered that low concentrations do not prevent functional interactions between proteins, and that the coexistence line between crystal and dilute solution in a protein phase diagram can occur at very low concentrations.[39] In our opinion, the negative design against crystallization is probably a mixture of direct and indirect effects.

In this article we have presented a different perspective by which to rationalize the crystallizability of proteins. Progress towards enhancing the success rate of crystallizing proteins will depend on unravelling the mechanisms by which nature achieves this negative design. We have highlighted several studies which show that random mutations enhance crystallizability. Mutagenesis programs have already led to important new insights into the nature of the functional interactions between proteins[26] and the key determinants of the propensity for amyloidogenic aggregation.[40] Similar systematic studies may provide an important means for understanding the mechanisms by which proteins are prevented from crystallizing. This would have the potential not only to provide further confirmation of our negative design hypothesis, but also to reveal residues and surface patterns that are key for the formation or prevention of crystal contacts.

We have already highlighted some interesting results that flag the potentially important role played by lsyine residues. Further, more detailed physical studies of the mechanisms by which lysine influences the protein-protein interactions would be desirable. For example, it would be interesting to see how the second virial coefficient, a measure of the strength of the generic attractions between proteins, changes with the mutation of surface lysine residues. Computer simulations could also potentially provide a more detailed atomistic picture of the conformations adopted by a surface lysine and how this changes with crystal contact formation.

Obtaining a better understanding of the mechanisms used to hinder crystallization would open up the possibility of finding ways to "turn off" these negative interactions, and so enhance a protein's crystallizability. The required changes to the surface properties could perhaps be achieved through mutations or the addition of appropriate precipitants. Furthermore, such advances in our understanding of protein crystallization could also potentially rationalize the effects of some of the precipitants currently used. At best, the effects of these precipitants are understood only in terms of their effect on average properties, such as the second virial coefficient. However, the mechanisms underlying some, e.g. polyethylene glycol, remain rather mysterious.

Finally, we note that only positive outcomes of protein crystallization experiments have traditionally been published. In our opinion, experiments where crystallizability is reduced rather than enhanced may also contain useful information about the mechanisms of negative design. Thinking in terms of this principle may help experimentalists decide when such "negative" results are nevertheless valuable.

To summarize, we have presented a perspective on protein crystallization whereby the difficulty crystallographers have in obtaining protein crystals is a consequence of evolutionary negative design against aggregation of native-state proteins. It really is the case that proteins do not want to crystallize because a protein that is prone to crystallization, or in fact any form of aggregation, is potentially deleterious to the cell. The mechanisms of this negative design are only very partially understood. But our main point is that understanding these mechanisms of *negative design* should provide fruitful insights that lead to *positive* advances in crystallizing globular proteins.

Acknowledgments

The authors are grateful to the Royal Society for financial support, and Luca Pellegrini for a careful read-

ing of the manuscript. The protein crystal image for the graphical table of contents was kindly provided by Allan D'Arcy, Morphochem, Switzerland.

* jpkd1@cam.ac.uk
† aal20@cam.ac.uk
‡ mv245@cam.ac.uk

[1] T. M. Bergfors, ed., *Protein Crystallization: Techniques, Strategies, and Tips* (International University Line, La Jolla, 1999).
[2] S. K. Burley, S. C. Almo, J. B. Bonanno, M. Capel, M. R. Chance, T. Gaasterland, D. Lin, A. Sali, F. W. Studier, and S. Swaminathan, Nat. Genet. **23**, 151 (1999).
[3] P. G. Vekilov and A. A. Chernov, Solid State Phys. **57**, 1 (2002).
[4] R. F. Service, Science **298**, 948 (2002).
[5] A. Yee, K. Pardee, D. Christendat, A. Savchenko, A. M. Edwards, and C. H. Arrowsmith, Acc. Chem. Res. **36**, 183 (2003).
[6] Glasses and gels most often form for systems where there is a network of strong bonds, eg SiO_2. By contrast, molecules with weak intermolecular interactions can relatively easily reorient to find the preferred orientation for the crystal.
[7] P. G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University, Ithaca, 1979).
[8] R. R. Kopito, Trends Cell Biol. **10**, 524 (2000).
[9] A. Pande, J. Pande, N. Asherie, A. Lomakin, O. Ogun, J. King, and G. B. Benedek, Proc. Natl. Acad. Sci. USA **98**, 6116 (2001).
[10] P. G. Vekilov, A. R. Feeling-Taylor, D. N. Petsev, O. Galkin, R. L. Nagel, and R. E. Hirsch, Biophys. J. **83**, 1147 (2002).
[11] A. McPherson, *Crystallization of biological macromolecules* (Cold Spring Harbour Laboratory Press, Cold Spring Harbour, 1999).
[12] H. W. Hellinga, Proc. Natl. Acad. Sci. USA **94**, 10015 (1997).
[13] W. F. DeGrado, C. M. Summa, V. Pavone, F. Nastri, and A. Lombardi, Annu. Rev. Biochem. **68**, 779 (1999).
[14] D. N. Bolon, Biochemistry **40**, 10047 (2001).
[15] J. S. Richardson and D. C. Richardson, Proc. Natl. Acad. Sci. USA **99**, 2754 (2002).
[16] W. Wang and M. H. Hecht, Proc. Natl. Acad. Sci. USA **99**, 2760 (2002).
[17] J. J. Havranek and P. B. Harbury, Nat. Struct. Biol. **10**, 45 (2003).
[18] J. M. Shifman and S. L. Mayo, Proc. Natl. Acad. Sci. USA **100**, 13274 (2003).
[19] R. J. Ellis and A. P. Minton, Nature **425**, 27 (2003).
[20] R. P. Sear, J. Chem. Phys. **120**, 998 (2004).
[21] R. P. Sear, unpublished.
[22] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, J. Mol. Evol. **56**, 243 (2003).
[23] H. E. McElroy, G. W. Sisson, W. E. Schoettlin, R. M. Aust, and J. E. Villafranca, J. Cryst. Growth **122**, 265 (1992).
[24] A. D'Arcy, M. Stihle, D. Kostrewa, and G. Dale, Acta Crystallogr. D **55**, 1623 (1999).
[25] G. E. Dale, C. Oefner, and A. D'Arcy, J. Struct. Biol. **142**, 88 (2003).
[26] A. A. Bogan and K. S. Thorn, J. Mol. Biol. **280**, 1 (1998).
[27] F. Baud and S. Karlin, Proc. Natl. Acad. Sci. USA **96**, 12494 (1999).
[28] L. Lins, A. Thomas, and R. Brasseur, Protein Sci. **12**, 1406 (2003).
[29] S. Dasgupta, G. H. Iyer, S. H. Bryant, C. E. Lawrence, and J. A. Bell, Proteins **28**, 494 (1997).
[30] G. H. Iyer, S. Dasgupta, and J. A. Bell, J. Cryst. Growth **217**, 429 (2000).
[31] L. Lo Conte, C. Chothia, and J. Janin, J. Mol. Biol. **285**, 2177 (1999).
[32] S. Jones and J. M. Thornton, Proc. Natl. Acad. Sci. USA **93**, 13 (1996).
[33] K. L. Longenecker, S. M. Garrard, P. J. Sheffield, and Z. S. Derewenda, Acta Crystallogr. D **57**, 679 (2001).
[34] K. L. Longenecker, M. E. Lewis, H. Chikumi, J. S. Gutkind, and Z. S. Derewenda, Structure **3**, 559 (2001).
[35] S. M. Garrard, K. L. Longenecker, M. E. Lewis, P. J. Sheffield, and Z. S. Derewenda, Protein Expres. Purif. **21**, 412 (2001).
[36] A. Mateja, Y. Devdjiev, D. Krowarsch, K. L. Longenecker, Z. Dauter, J. Otlewski, and Z. S. Derewenda, Acta Crystallogr. D **58**, 1983 (2002).
[37] E. Schnepf, N. Crickmore, J. Van Rie, D. Lereclus, J. Baum, J. Feitelson, D. R. Zeigler, and D. D. H., Microbiol. Mol. Biol. Rev. **62**, 775 (1998).
[38] K. M. Smith, *Virus-Insect Relationships* (Longman, London, 1976).
[39] M. Muschol and F. Rosenberger, J. Chem. Phys. **64**, 1953 (1997).
[40] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, and C. M. Dobson, Nature **424**, 805 (2003).