Detecting native-like properties in combinatorial libraries of *de novo* proteins

Sushmita Roy, Kimberly J Helmer and Michael H Hecht

Background: Combinatorial methods based on binary patterning of polar and nonpolar residues have been used to generate large libraries of *de novo* α-helical proteins. Within such libraries, the ability to find structures that resemble natural proteins requires a rapid method to sort through large collections of proteins and detect those possessing 'native-like' features. The current paper presents such a method and applies it to an initial collection of *de novo* proteins.

Results: We present a method to identify proteins with native-like properties from libraries of *de novo* sequences expressed *in vivo.* A novel 'rapid prep' freeze/thaw procedure was used to prepare samples; chromatographic purification was not required. The semi-crude samples were analyzed for native-like features by one-dimensional ^{1}H NMR spectroscopy. Using this method, we demonstrate that native-like features can readily be observed for several proteins among a collection of sequences designed by binary patterning of polar and nonpolar amino acids.

Conclusions: Native-like properties can be detected using a method that requires neither isotopic enrichment nor chromatographic purification. The method is inexpensive, rapid, and suitable for parallel processing. It can therefore be employed to screen for native-like properties among large collections of *de novo* sequences. Using this method, we demonstrate that although the binary code strategy does not explicitly design tertiary packing, it can nonetheless generate proteins that possess native-like properties. The use of combinatorial methods to produce large collections of proteins coupled with the availability of a rapid assay for detecting native-like properties will facilitate the design and isolation of novel proteins with desirable properties.

Address: Department of Chemistry, Princeton University, Princeton, NJ 08544-1009, USA.

Correspondence: Michael H Hecht
e-mail: hecht@berthaw.princeton.edu

Key words: binary code proteins, *de novo* proteins, native-like properties, NMR screen, protein design

Received: 11 Nov 1996
Revisions requested: 02 Dec 1996
Revisions received: 23 Dec 1996
Accepted: 13 Jan 1997

Published: 29 Jan 1997
Electronic identifier: 1359-0278-002-00089

Folding & Design 29 Jan 1997, 2:89–92

© Current Biology Ltd ISSN 1359-0278

Introduction

Combinatorial methods for synthesizing novel genes have made it possible to construct and express large libraries of *de novo* protein sequences [1–4]. The vast number of sequences that can be generated and the ease with which the resulting proteins can be produced heighten the need for methods to screen these collections for sequences possessing desired properties.

We have previously described a combinatorial strategy for producing libraries of *de novo* α-helical proteins [2]. Our strategy is based on the premise that the sequence periodicity of polar and nonpolar amino acids is a major determinant of secondary structure, and that amphiphilic secondary structures will readily collapse into compact globular tertiary folds [5,6]. Using combinatorial genetic methods, we produced a library of sequences in which the sequence positions of hydrophobic and hydrophilic residues were specified explicitly, but the precise identities of these residues were varied extensively [2]. Thus, the design was based upon a 'binary code' that specified only whether a given residue would be polar or nonpolar. Using degenerate codons for generic polar and nonpolar residues, a collection was constructed that could, in principle, contain 5×10^{41} different amino acid sequences. Characterization of a subset of these proteins indicated that the majority formed structures that were soluble, stable, and α-helical.

The combinatorial underpinnings of the binary code strategy preclude explicit design of specific tertiary interactions. Therefore it was unlikely that all of the *de novo* α-helical proteins would fold into unique native-like tertiary structures. Nonetheless, the ability of the binary code strategy to generate large collections of soluble α-helical proteins led us to investigate whether proteins with native-like properties might be found in our collection.

Native proteins possess a number of biophysical properties that distinguish them from molten globule folding intermediates. These properties include a well defined oligomeric state, chemical shift dispersion in the NMR spectrum, protection from amide proton exchange, and a

Figure 1



Coomassie-stained SDS-polyacrylamide gel of *de novo* protein G. Left, whole cells. Center, NMR sample from the rapid prep procedure. Right, chromatographically purified protein.

denaturation profile that is both cooperative and occurs with a significant positive ΔCp. Proteins chosen from a combinatorial library of *de novo* sequences might possess all, some, or none of these properties. Characterization of these properties for a large collection of *de novo* proteins is a daunting task — particularly if each protein must be purified to homogeneity prior to analysis. Therefore, we wished to develop a method that would facilitate rapid detection of the most native-like members of the collection, so that these individual proteins could be singled out for purification and detailed characterization.

Ideally, such a method would be rapid, inexpensive, capable of analyzing semi-pure samples that do not require arduous purification, amenable to processing many samples in parallel, and, most importantly, able to discriminate between collapsed 'molten' structures and structures with native-like features. Here, we describe a method that meets these five criteria. We apply this method to our initial collection of *de novo* α-helical proteins [2] and show that several proteins with native-like properties can readily be distinguished from those with molten globule like behavior.

Results and discussion

The 29 proteins in our initial collection [2] of binary code sequences were analyzed. The proteins were not chromatographically purified. Semi-pure samples of highly expressed protein were prepared from bacterial cells by a modification of a freeze/thaw protocol described previously [7]. Freezing and thawing disrupts the integrity of the cell envelope, thereby allowing small recombinant proteins to leak out. Because the cells are not actually lyzed, nucleic acids and the vast majority of *Escherichia coli* proteins are left behind in the unlyzed cells, which are readily removed by centrifugation. Although the freeze/thaw treatment allows some endogenous bacterial proteins to leak out also, most of these can be removed by acid precipitation. The resulting 'rapid prep' samples contain the *de novo* protein as the dominant product (see Fig. 1). Because the rapid prep procedure requires no chromatographic separations, it can be adjusted to any scale and can be performed on many samples in parallel.

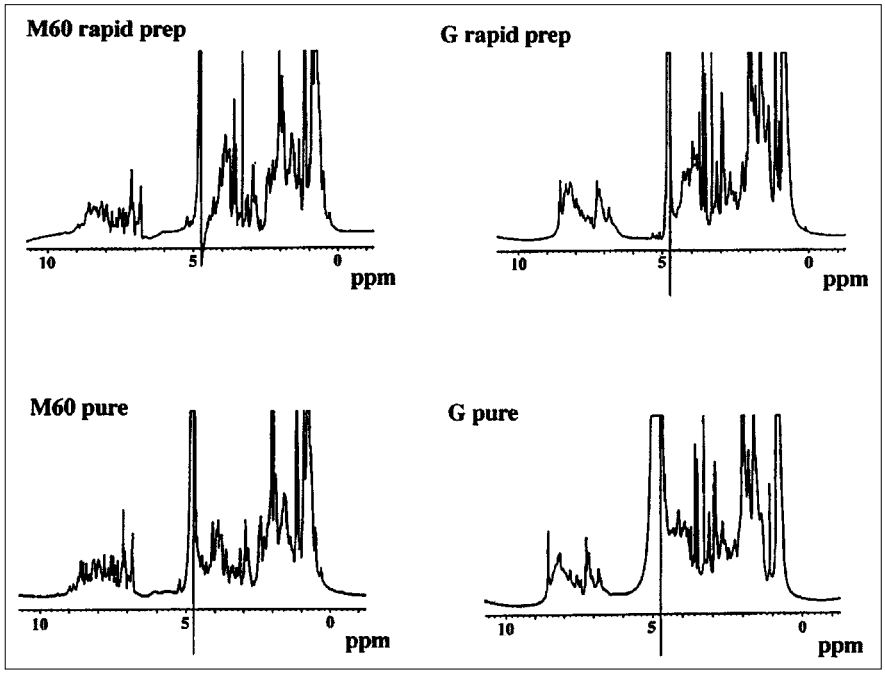
NMR studies of proteins are conventionally carried out on highly purified isotopically labeled samples [8,9]. However, the speed, efficiency, and high yield of the rapid prep procedure led us to investigate whether semi-pure unlabeled samples might be suitable to identify native-like features in the ^{1}H NMR spectra of the *de novo* proteins. Because the contaminating cellular proteins that remain after the freeze/thaw and acid precipitation procedures are both larger and less abundant than our *de novo* proteins, we reasoned that the NMR peaks resulting from these contaminants would be broadened and too small to rise above baseline.

Figure 2 compares the 1D ^{1}H NMR spectra of proteins prepared by the rapid prep procedure to the spectra of the same proteins purified to >98% (see Fig. 1). For both proteins, the spectra obtained from the rapid prep samples are virtually identical to those obtained from purified protein. The impurities remaining in the rapid prep samples do not interfere with the ^{1}H resonances from the overexpressed *de novo* protein.

The spectrum of protein G shows poor chemical shift dispersion and no evidence of native-like features. Its 1D ^{1}H NMR spectrum suggests that the sequence of protein G does not fold into a unique well defined structure. In contrast, the spectrum of protein M60 indeed displays native-like features. Its spectrum is well dispersed in both the amide (6.5–9.5 ppm) and the methyl regions (upfield of 3 ppm). Specific peaks suggestive of native-like structure include the ring-current shifted methyl peaks at 0.5, 0.3, and –0.4 ppm, as well as the amide proton peak at 9.4 ppm.

As the semi-pure samples are sufficient to distinguish sequences with good chemical shift dispersion from those with poor dispersion, we prepared rapid prep samples for

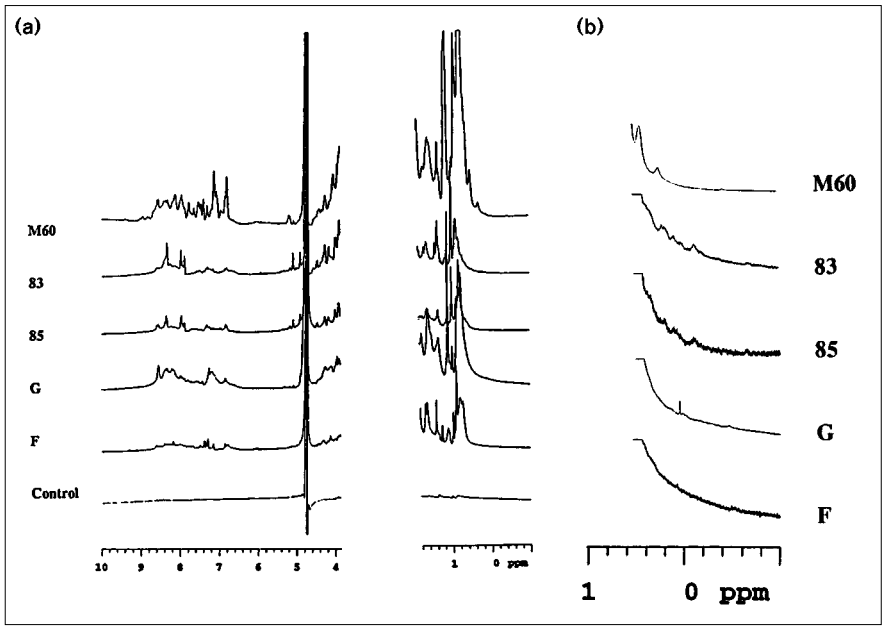
Figure 2



Comparison of the 1D ^{1}H NMR spectra of rapid prep samples to the spectra of chromatographically purified proteins M60 and G. The spectra obtained from rapid prep samples are identical to spectra of the same proteins purified to >98%. In the spectra of protein M60, native-like features are observed. In particular, note the well dispersed amide region (6.5–9.5 ppm) and the upfield shifted methyl peaks at 0.5, 0.3, and –0.4 ppm. In the spectra of protein G, these native-like features are absent. (The sharp spike at ~0.05 ppm in the rapid prep spectrum of protein G results from a small molecule contaminant.)

all 29 sequences in our initial binary code collection and measured their ^{1}H NMR spectra. As shown in Figure 3, a range of behaviors can be observed. Some of the proteins yield spectra with poor chemical shift dispersion (e.g. G and F), others have moderate dispersion (e.g. 83 and 85), and some show significant dispersion (e.g. M60). The proteins with most dispersion are likely to be the most native-like. Indeed, initial characterizations indicate a good correlation between proteins of the collection with good chemical shift dispersion and those which exhibit cooperative thermal denaturation profiles (S Roy, MH Hecht, unpublished data).

Figure 3



1D ^{1}H NMR spectra of *de novo* proteins. (a) The 1D ^{1}H NMR spectra of several binary code proteins obtained by the rapid prep procedure. The control spectrum (bottom) was measured for a sample prepared from the same strain of *E. coli* harboring the control expression plasmid (pET-3D). Although its vertical scale is identical to the other samples, the control sample shows no significant resonances. Thus, the impurities remaining in the rapid prep samples do not interfere with the ^{1}H resonances from the overexpressed *de novo* protein. (b) Enlargements of the region where upfield shifted methyl peaks (characteristic of native proteins) would be observed. Proteins M60, 83, and 85 clearly show several upfield shifted resonances, whereas proteins G and F do not. (The sharp spike at ~0.05 ppm in the spectrum of protein G results from a small molecule contaminant.)

Our method for detecting native-like properties in a library of *de novo* proteins can be compared to a method reported by Gronenborn and Clore [10] to screen for structural integrity among mutants of natural proteins. They describe a procedure based on ^{15}N labeling and ^{1}H–^{15}N heteronuclear single quantum coherent spectroscopy of recombinant proteins in total cell lysates. While our procedure does not provide as detailed structural information, it is well suited as an initial screen for native-like features in combinatorial libraries of *de novo* proteins. The freeze/thaw and acid precipitation procedure can be carried out on several dozen samples in parallel in less than 2 h. Analysis of these samples does not require isotopic labeling. ^{1}H NMR spectra can be acquired very rapidly on standard high field NMR instruments and do not require the additional hardware or pulse sequences necessary for multidimensional experiments. Finally, since all samples have identical volume, are in the same buffer and have similar impurities, the maximum time spent on shimming samples was 1 min. In some cases, four transients were sufficient to distinguish native-like properties.

In conclusion, we have demonstrated that 1D ^{1}H NMR spectra of overexpressed protein samples prepared by the rapid prep procedure are virtually identical to spectra of the same proteins purified chromatographically. These spectra can be used to distinguish proteins with native-like properties from among libraries of *de novo* sequences. Using this method, we have found several proteins in our collection of binary code sequences that yield spectra with reasonable chemical shift dispersion. This indicates that although the binary code strategy does not explicitly design tertiary packing, it can nonetheless generate proteins that possess native-like properties.

Materials and methods

Cell growth

250 ml cultures of transformed cells were grown in 2XYT (100 μg ml^{-1} ampicillin) at 37°C to an OD_{600} between 0.7 and 1. The culture was induced with 100 μg ml^{-1} IPTG for 3 h and cells were harvested by centrifugation at 6000 × g for 10 min at 4°C.

Rapid prep procedure

Harvested cell pellets were subjected to three cycles of freezing (dry-ice/ethanol bath) and thawing (10°C water bath) for 10 min each. Following the third cycle, 10 ml of 100 mM $MgCl_2$ was added. Samples were kept on ice for 1 h with gentle inversion to resuspend cells. Samples were then centrifuged at 6000 × g for 10 min at 4°C. The supernatant containing the protein of interest was carefully decanted. The freeze/thaw supernatant was acidified by addition of 1 M sodium acetate/acetic acid buffer (pH 4.0) to a final concentration of 50 mM. Samples were kept on ice for 15 min and then centrifuged at 6000 × g for 15 min at 4°C to precipitate contaminants. Providing that acid precipitation of the cellular contaminants is carried out in the presence of reasonable concentrations of salt (>100 mM), we have found that small *de novo* proteins typically remain in solution. The supernatant was concentrated in a Centricon-3 concentrator (Amicon) and exchanged into 50 mM CD_3COOD/NaOH, pH 4.6 and 200 mM NaCl. Final sample volume was made to 400 μl and 10% D_2O was added. Protein concentrations typically ranged between 0.3–0.8 mM.

For highly expressed proteins, the 1D ^{1}H NMR of the freeze/thaw supernatant sample (i.e. prior to removal of contaminants by acid precipitation) also yields high quality spectra. This technique may therefore also be useful for proteins that do not survive acid precipitation.

NMR spectroscopy

1D ^{1}H NMR spectra were recorded at 25°C on a Varian Inova 500 MHz spectrometer. Water was suppressed using a presaturation pulse. Spectra shown in Figure 2 were acquired with 128 transients and total acquisition time of ~5 min.

Acknowledgements

We thank Jim Beasley, Judith Swan, Gayathri Ratnaswamy, and Fred Hughson for helpful suggestions. This work was supported by a grant from the New Jersey Commission on Science and Technology, and by a Beckman Young Investigator award (to MH Hecht).

References

1. Mandecki, W. (1990). A method for construction of long randomized open reading frames and polypeptides. *Protein Eng.* 3, 221–226.
2. Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M. & Hecht, M.H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science* 262, 1680–1685.
3. Davidson, A.R. & Sauer, R.T. (1994). Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. USA* 91, 2146–2150.
4. Davidson, A.R., Lumb, K.J. & Sauer, R.T. (1995). Cooperatively folded proteins in random sequence libraries. *Nat. Struct. Biol.* 2, 856–864.
5. Xiong, H., Buckwalter, B.L., Shieh, H.-M. & Hecht, M.H. (1995). Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc. Natl. Acad. Sci. USA* 92, 6349–6353.
6. West, M.W. & Hecht, M.H. (1995). Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci.* 4, 2032–2039.
7. Johnson, B.H. & Hecht, M.H. (1994). Recombinant proteins can be isolated from *E. coli* cells by repeated cycles of freezing and thawing. *Biotechnology* 12, 1357–1360.
8. Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids.* John Wiley & Sons, New York.
9. Clore, G.M. & Gronenborn, A.M. (1991). Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy. *Science* 252, 1390–1399.
10. Gronenborn, A.M. & Clore G.M. (1996). Rapid screening for structural integrity of expressed proteins by heteronuclear NMR spectroscopy. *Protein Sci.* 5, 174–177.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published via the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad.htm – for further information, see the explanation on the contents page.