

# Conserved signature proposed for folding in the lipocalin superfamily

Lesley H. Greene<sup>a,b,\*</sup>, Daizo Hamada<sup>b,1</sup>, Stephen J. Eyles<sup>c,2</sup>, Keith Brew<sup>a,3</sup>

<sup>a</sup>Department of Biochemistry and Molecular Biology, University of Miami School of Medicine, Miami, FL 33101, USA

<sup>b</sup>Oxford Centre for Molecular Sciences, Central Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QH, UK

<sup>c</sup>Department of Chemistry, University of Massachusetts at Amherst, Amherst, MA 01003, USA

Received 22 April 2003; revised 1 July 2003; accepted 17 July 2003

First published online 11 September 2003

Edited by Thomas L. James

**Abstract** We systematically identify a group of evolutionarily conserved residues proposed for folding in a model  $\beta$ -barrel superfamily, the lipocalins. The nature of conservation at the structural level is defined and we show that the conserved residues are involved in a network of interactions that form the core of the fold. Exploratory kinetic studies are conducted with a model superfamily member, human serum retinol-binding protein, to examine their role. The present results, coupled with key experimental studies conducted with another lipocalin  $\beta$ -lactoglobulin, suggest that the evolutionarily conserved regions fold on a faster folding time-scale than the non-conserved regions. © 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.

**Key words:** Evolution; Protein folding; Conserved residue; Lipocalin superfamily; Retinol-binding protein;  $\beta$ -Lactoglobulin

## 1. Introduction

The disparity between the rapid folding of proteins and the astronomical time required for a polypeptide chain to find its thermodynamically stable, native state by random search indicates that the mechanistic process is non-random [1]. The elucidation of underlying sequence and structural determinants that bias the energetics of the folding process should provide a common conceptual framework to better understand the rapid folding of protein domains. It has been proposed that the search process is guided by the formation of a folding nucleus that may have been selected for and maintained throughout the evolution of protein folds to enable correct folding in relevant biological time [2–6]. Although conserved sequences in homologous proteins have been generally thought to be regions that are necessary for function [7], the link between sequence conservation and protein folding has recently gained considerable attention [8–16].

We identify a group of conserved residues in the lipocalin superfamily that we propose were selected for and maintained

during evolution to direct folding [2,4,5]. This superfamily was selected for analysis based on the following criteria [2,4,5]. (1) They are functionally diverse [17] which suggests that conservation of specific residues for function does not limit their sequence variability. (2) They are very divergent in sequence with many members sharing under 25% sequence identity yet share a common fold which consists of an eight-stranded  $\beta$ -barrel with a small C-terminal helix and an average chain length of 175 residues [18,19]. The low sequence identity and similar fold suggest that a minimalist set of common residues forming a conserved sequence and structural signature for this  $\beta$ -barrel topology may be identifiable. (3) They are an ancient group with members found in both eukaryotes and prokaryotes [20].

To explore the proposed role of conserved residues in folding using the human serum retinol-binding protein (RBP) as a model for the lipocalins we first aim to determine if it is possible to distinguish between the folding of conserved versus non-conserved regions of the protein. Our initial kinetic studies for RBP and a variant suggest that there are two significant stages of folding: collapse of the core and formation of tertiary structure. They are determined by independently monitoring the formation of both regions with biophysical techniques and probes that provide structural resolution of the kinetic process. We also show a correlation between our conservation data and key kinetic studies conducted previously with another lipocalin,  $\beta$ -lactoglobulin [21], that provides support for the early role of conserved residues for folding in this superfamily.

## 2. Materials and methods

### 2.1. Sequence and structural studies

A group of 32 lipocalins were chosen to provide a sample with maximum diversity of sequence and function. The construction of a superfamily multiple sequence alignment is described elsewhere [2,5]. Sequence conservation,  $C(i)$ , is defined as,

$$C(i) = 1 - S(i)/\ln(m) \quad (1)$$

where  $S(i)$  is the sequence entropy of residue  $i$  calculated for each position in the alignment where there were 2 or fewer deletions and  $m$  represents the 20 amino acids used in this analysis [22]. Conservation,  $C(i)$ , is a parameter that can vary from 0 where all 20 amino acid residues are equally represented to 1.0 where there is absolute conservation of one amino acid residue. The alignment was also analyzed for the average side chain hydrophobicity at each position using a defined hydrophathy scale [23].

### 2.2. Proteins and materials

The recombinant wild-type RBP and Trp24 only variant used in these studies were prepared in accordance with previously established protocols [24]. Protein concentrations were determined by absorption

\*Corresponding author. Fax: (44)-1865-275905.

E-mail address: lesley.greene@bioch.ox.ac.uk (L.H. Greene).

<sup>1</sup> Present address: Department of Developmental Infectious Diseases, Research Institute and Osaka Medical Center for Maternal and Child Health, 840 Murodo-cho, Izumi, Osaka 594-1011, Japan.

<sup>2</sup> Present address: Department of Polymer Science and Engineering, University of Massachusetts, Amherst, MA 01003, USA.

<sup>3</sup> Present address: Department of Biomedical Sciences, Florida Atlantic University, 777 Glades Road, P.O. Box 3091, Boca Raton, FL 33431, USA.

at 280 nm using calculated extinction coefficients. The precise concentrations of guanidine hydrochloride (GndHCl) in sample solutions were determined by refractive index at 25°C.

### 2.3. Folding studies

Stopped-flow fluorescence data were collected with a Bio-Logic SFM-4 stopped-flow module attached to a PTI QM-1 fluorimeter (Photon Technologies, Inc) with a slit width of 1 nm and a FC-08 cuvette. Protein concentration prior to dilution was 32  $\mu$ M and denatured in 4 M GndHCl and 5 mM sodium phosphate buffer (pH 7.0) for 2 h before initiating refolding by 11-fold dilution in 5 mM sodium phosphate buffer (pH 7.0). Baselines for the native and denatured states using a titration with GndHCl and monitored by fluorescence spectroscopy were determined previously [24]. Stopped-flow near-UV circular dichroism (CD) refolding studies were performed with a Bio-Logic SFM-4 stopped-flow CD system with a slit width of 2 nm and a TC100-15 cuvette. The refolding conditions are exactly the same as in the fluorescence experiment except the starting concentration is 48  $\mu$ M. The traces for the stopped-flow studies were averaged from separate shots performed under identical conditions at 20°C. Kinetic traces for both studies were well fit to either single or double exponential functions using the scientific curve-fitting program, Sigma Plot® (ver. 4.1). Residuals (simple difference between the data values and predicted fluorescence or CD from the curve fit) were plotted against time to determine the quality of fit. The small magnitude and approximately equal distribution centered around zero indicate that the mathematical model is appropriate. Stopped-flow far-UV CD studies to determine the time-scale of secondary structure formation

are not possible because the native far-UV CD spectrum of RBP contains contributions from aromatic residues in the 220–235 nm region and the peptide CD signal cannot be clearly separated [24,25].

## 3. Results and discussion

### 3.1. Sequence and structural studies

Distributions of conservation values ( $C(i)$ ) for all aligned sites (derived from an analysis of the multiple sequence alignment) are highly skewed or bimodal (Fig. 1A). The distribution identifies a major group of sites with a reasonably Gaussian distribution which we subdivide into two groups: one with a low level of conservation ( $0 < C(i) < 0.35$ ) and a group with intermediate scores ( $0.35 \leq C(i) < 0.45$ ). It also identifies a small group of highly conserved sites ( $C(i) \geq 0.45$ ). The conservation profile reveals that the intermediate and highly conserved sites are grouped in five discrete regions of sequence. We designate the regions inclusive of the conserved residues as ECRs (evolutionarily conserved regions) I–V (Fig. 1A). Two isolated sites do not fit this pattern of conservation and correspond in many lipocalins to a pair of disulfide bonded cysteines which links the C-terminus to  $\beta$ -strand C. Future studies will be required to determine if they are conserved for

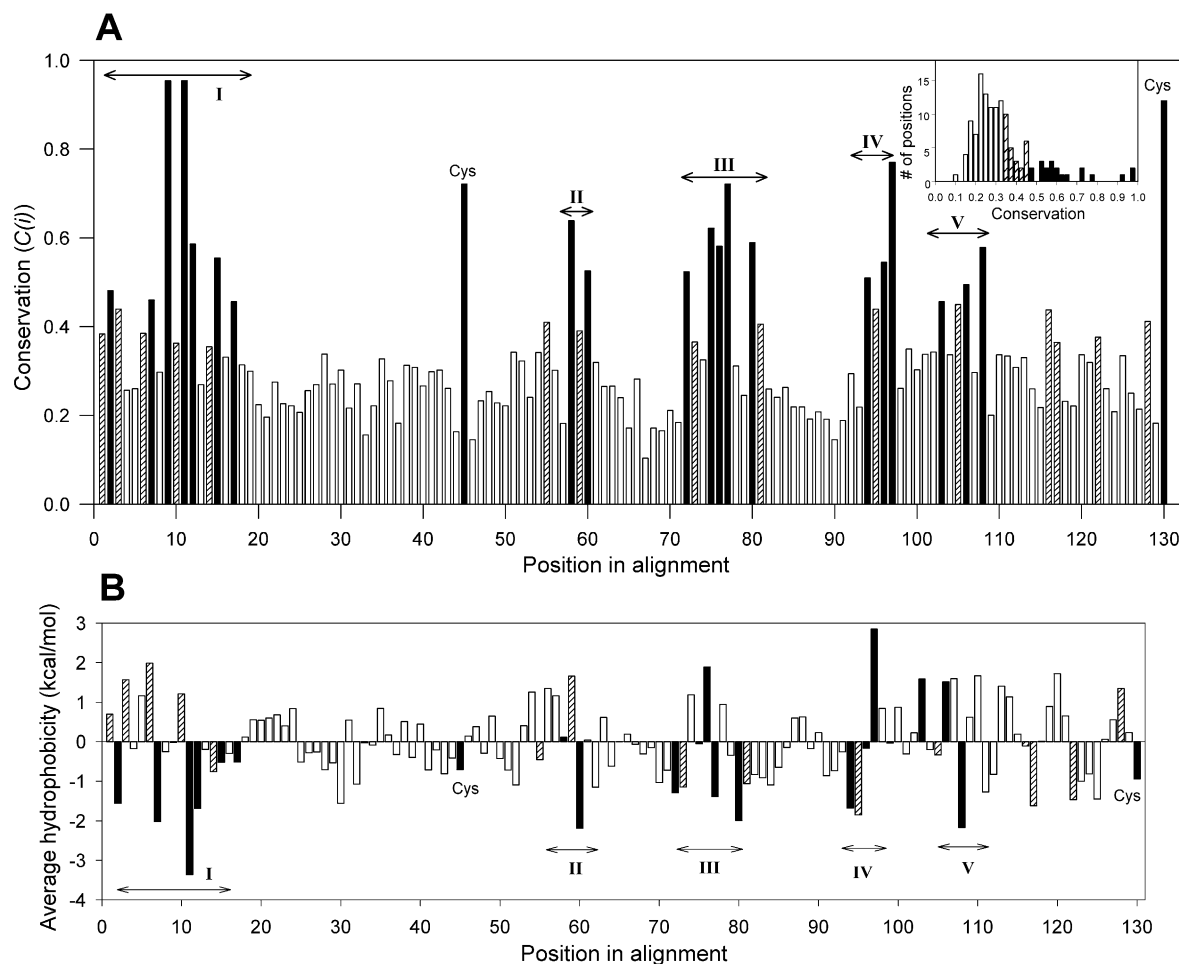


Fig. 1. Sequence conservation analysis of the lipocalins. A: The insert shows the distribution of conservation ( $C(i)$ ) scores grouped in 0.02 increments. In the graphs the three levels of conservation are represented by white bars  $C(i) < 0.35$ , patterned bars  $0.35 \leq C(i) < 0.45$  and black bars  $C(i) \geq 0.45$ . The ECRs are annotated with roman numerals. Consensus sequences for residues with ( $C(i)$ ) scores  $\geq 0.45$ , which are the black bars: ECR I ( $\beta$ -strand A including a  $\beta$ -bulge) xFxxxxF/YxGxWYxxAxA; ECR II (D/E  $\beta$ -bend) G/AxY/Y; ECR III ( $\beta$ -strand F and F/G bend) VxxTDYxxF/Yx; ECR IV ( $\beta$ -strand H) LxG/SR; ECR V (helix) ExxExF. B: Average hydrophobicity profile analysis in the lipocalins.

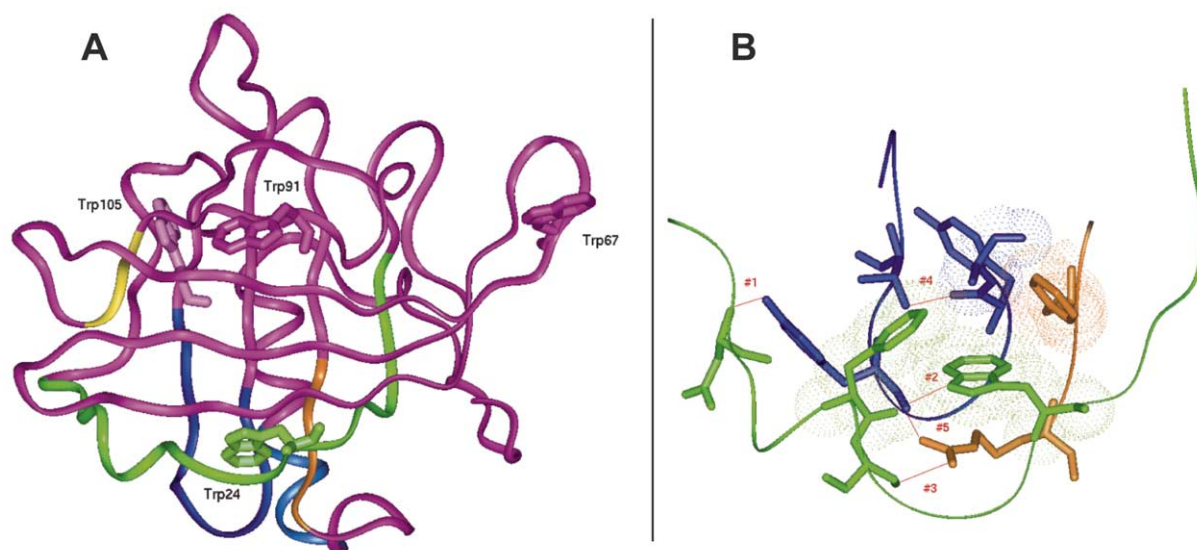


Fig. 2. Structural analysis of RBP. A: Ribbon model of human serum RBP (pdb code 1rbp). Relative locations of the four tryptophans are shown and labelled accordingly. Conserved structural features are color-coded to represent ECR I (green: residues 14–30), ECR II (yellow: residues 84–86), ECR III (blue: residues 106–115), ECR IV (orange: residues 136–139) and ECR V (light blue: residues 146–151). Non-conserved regions are colored in purple. B: View of RBP showing a hydrophobic cluster between Phe20, Trp24, Ala115 and Phe137; an amine-aromatic interaction between Trp24 and Arg139; tertiary hydrogen bonds are labeled, #1 Asp16(N)-Tyr111(OH), #2 Phe20(CO)-Trp24(NE1), #3 Ser21(CO)-Arg139(NH1), #4 Thr109(OG1)-Tyr114(CO), #5 Tyr111(CO)-Arg139(NH2). 1RBP is visualized with InsightII, Version 98 (Accelrys, San Diego, CA, USA).

stability as we would predict or have a role in folding as well. Some residues in three of the ECRs were noted in an earlier and less systematic analysis, of a smaller group of lipocalins. These conserved residues were postulated to be required for a common physiological function [26], a perspective that is in-

consistent with their functional diversity and location in the structure.

Examination of the locations of the ECRs within the lipocalin structures reveal that although located in distinct regions of the polypeptide chain the ECRs are grouped together in the

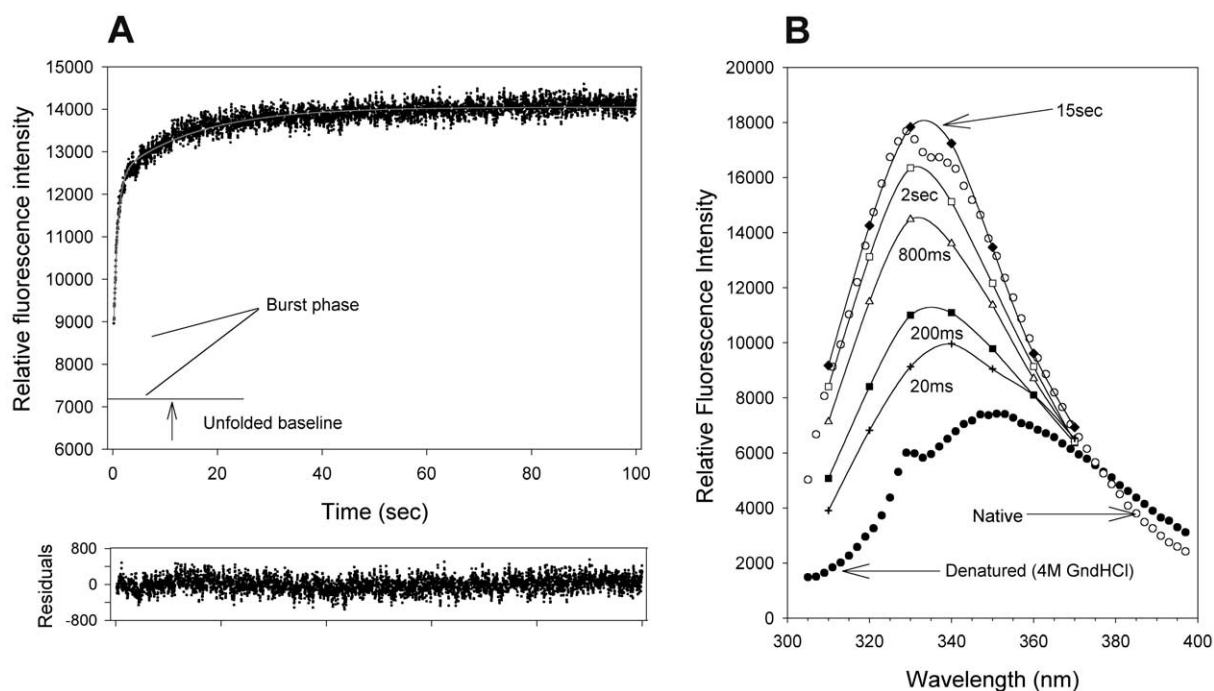


Fig. 3. Formation of the conserved core probed by the refolding of the Trp24 only variant. A: Refolding trace of the Trp24 only variant using stopped-flow fluorescence spectroscopy is shown. The  $\lambda_{\text{ex}} = 295$  nm and the change in fluorescence is monitored at  $\lambda_{\text{em}} = 350$  nm, both of which are selective for tryptophan. The rate constants and amplitudes are:  $k_1 = 1.27 \pm 0.03 \text{ s}^{-1}$ ,  $A_1 = -4390 \pm 82$  and  $k_2 = 0.066 \pm 0.001 \text{ s}^{-1}$ ,  $A_2 = -1650 \pm 18$ . B: Reconstruction of the folding for the Trp24 only variant. The emission wavelengths were monitored between 310 and 370 nm with  $\lambda_{\text{ex}} = 295$  nm. The spectra are plotted by wavelength as a function of fluorescence intensity at the selected time-points shown.

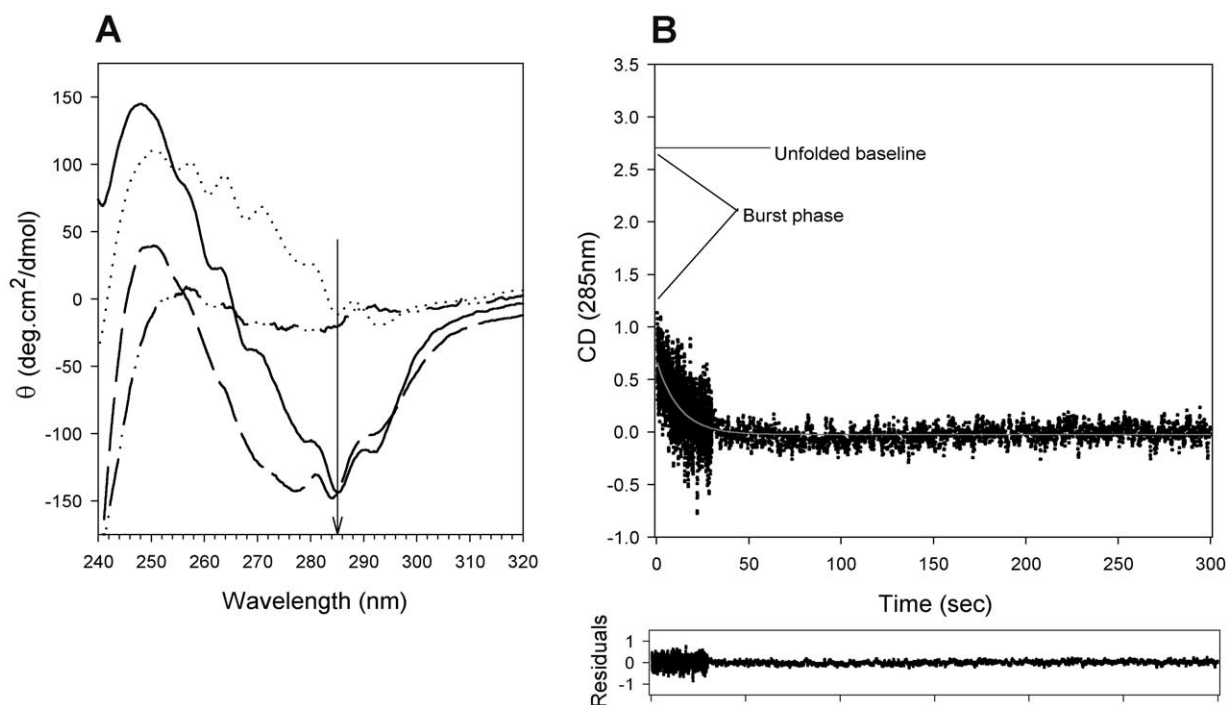


Fig. 4. Formation of wild-type RBP native structure. A: Near-UV(240–320 nm) CD spectrum of RBP and variants. Wild-type RBP – native state, pH 7.4 (solid line); unfolded state, 6 M GndHCl, pH 7.0 (dash-dotted line); W67L/W91H/W105F variant (dotted line); W24L variant (dashed line). B: Refolding of RBP as monitored by stopped-flow near-UV CD. CD spectra were monitored at 285 nm following an 11-fold dilution from 4 M GndHCl. The kinetic traces were averaged and fit to a single exponential. The rate constant and amplitude is  $k_1 = 0.082 \pm 0.002 \text{ s}^{-1}$  and  $A_1 = 0.71 \pm 0.01$ .

3-D structures (Fig. 2A). In lipocalins of known structure represented by RBP, they form a mixed polar/non-polar core that closes the base of the barrel. The core, which we term the ‘fold-determining core’ for its proposed role in folding and stability, appears to be particularly stabilized by a network of long-range interactions between a section of  $\beta$ -strand A (ECR I), the turn between  $\beta$ -strands F and G (ECR III) and the C-terminal end of  $\beta$ -strand H (ECR IV) and within ECRs I and III (Fig. 2B). Interestingly, the average hydrophobicity profile generated from the lipocalin alignment reveals that the most persistently hydrophobic sites are within the ECRs (Fig. 1B). This suggests that the hydrophobic effect will promote the concentration of ECRs in the collapsed state, enhancing the rate of formation of cooperative polar and non-polar long-range interactions that form the core.

### 3.2. Folding studies

RBP was characterized with respect to structure, function and stability as described previously where it was also shown that RBP folds and unfolds reversibly by chemical denaturation which is an essential condition for folding kinetics studies [24]. Initial stopped-flow fluorescent studies on RBP indicated that the folding process is multiphasic and therefore a suitable experimental model to look for the proposed differences in rates between the conserved and non-conserved regions identified by our conservation analysis (L. Greene et al., unpublished data). To initially test the hypothesis that conserved residues guide folding by becoming structured on a faster folding time-scale than the non-conserved regions, an exploratory kinetics experiment was designed. The first part of the experiment involved the use of stopped-flow fluorescence spectroscopy. There are four endogenous tryptophans in RBP that

can serve as fluorescent probes. One is highly conserved (Trp24) in comparison to the other three (Trp67, Trp91, Trp105) and they are all in distinct locations within the protein and can therefore monitor different regions of the protein (Fig. 2A).

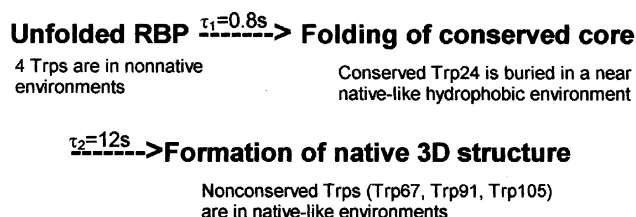
A Trp24 only variant was constructed to provide a fluorescence probe that would monitor the folding of the conserved core (Fig. 2A). Characterization of this protein showed that it has native-like stability [24] and is functional [5]. The Trp24 only variant shows biphasic kinetics during refolding from the denatured state (Fig. 3A). The relaxation times for the first phase and second phase are 0.79s and 15.2 s, respectively. The majority of the fluorescence change occurs during the first phase (73%), which indicates that the conserved Trp24 is sequestered from solvent early in folding. Monitoring the emission spectra at 10 nm intervals between 310 and 370 nm and reconstructing the spectra reveals that Trp24 is in a near native-like hydrophobic environment during the first phase (Fig. 3B). This interpretation is based on fluorescence theory, whereby the native state has higher intensity than the denatured state and the peak intensity is blue-shifted as the fluorescent probe moves to an internal hydrophobic environment [27].

The next step was to determine if we could monitor the folding of the non-conserved regions for comparison. Near-UV CD is an ideal technique for examining the acquisition of native structure and has been used successfully to study the folding of several proteins [28–30]. This is accomplished by monitoring the development of native-like tertiary packing of predominantly the aromatic residue tryptophan with respect to time. We chose this method to dissect the tryptophan signals over using a variant with a single fluorescent Trp probe

in the non-conserved regions because stability studies revealed that substitution of Trp24 to Tyr, Phe or Leu significantly destabilized the native state [24].

Equilibrium near-UV and far-UV CD studies of RBP and Trp variants were found not to globally disrupt the native tertiary structure of RBP [24]. The CD spectra of the mutants allow for the deconvolution of contributions of individual and groups of tryptophans to the CD spectra of RBP thereby enabling the design and interpretation of stopped-flow CD studies. The substitution to the conserved tryptophan has a different effect on the CD spectrum in comparison to the other tryptophans, with only a small loss in molar ellipticity at 285 nm (Fig. 4A). At 285 nm the CD signal is dominated by the three non-conserved Trps and not by the conserved Trp24 (Fig. 4A). Thus the non-conserved Trps are good probes for native state formation. Analysis of the stopped-flow near-UV CD studies with RBP indicates that the rate of formation of the native structure occurs in a single phase with a relaxation time of 12.2 s (Fig. 4B).

These studies indicate that the conserved core probed by Trp24 using stopped-flow fluorescence spectroscopy is buried in a hydrophobic environment at the early stages of folding in contrast to the later formation of the native state as monitored by stopped-flow CD. In summary:



#### 4. Conclusion

In this paper we systematically detailed the nature of sequence and structural conservation proposed for folding in the lipocalin superfamily. We hypothesized that the conserved core would fold first. We designed experiments to test this hypothesis and provided initial experimental support using RBP as an experimental model. Future work using other biophysical techniques such as NMR spectroscopy is required to further test whether Trp24 and other conserved residues form their native contacts early and in a straightforward manner, which is expected of a folding nucleus.

Support for the proposed role of conserved residues for folding in the lipocalins comes from a key kinetic hydrogen-deuterium protection experiment, with  $\beta$ -lactoglobulin, the only other lipocalin to have been characterized kinetically [21,31–33]. The experimental result of Kuwata et al. clearly indicates that  $\beta$ -strands F, G, H and the C-terminal helix are protected early [21]. These regions correspond to ECR III, ECR IV and ECR V based on our sequence and structural conservation analysis (Fig. 2A). The results of this kinetic study are in part reinforced by studies of  $\beta$ -lactoglobulin peptides which show that there are native-like hydrophobic interactions involving  $\beta$ -strands G and H which comprise part of ECR III and ECR IV [34].

On the other hand, the experimental results of Kuwata et al. suggest that weak protection is found in a segment span-

ning residues 12–21 in  $\beta$ -strand A, which is consistent with formation of marginally stable non-native  $\alpha$ -helix near the N-terminus at the early stages of folding [21,31]. This segment is part of ECR I. Peptide studies of  $\beta$ -lactoglobulin which include  $\beta$ -strand A also indicate that this region has an intrinsic preference to adopt a helical conformation and may contribute to the formation of this transient non-native helix [35]. There is no experimental evidence at present for this characteristic in RBP. Interestingly, the results from four secondary structure prediction programs (PHD [36], nnPredict [37], GorIV and HNN [38]) suggest that  $\beta$ -lactoglobulin has a higher intrinsic helical preference than RBP in  $\beta$ -strand A, but is not confirmatory because structure prediction programs are not completely accurate (data not shown). Ultimately it will be interesting to determine if variations may exist regarding the importance of each ECR between these two proteins. Analogous kinetic studies, particularly at atomic resolution are also needed for other members of lipocalin superfamily to further test this hypothesis.

**Acknowledgements:** We are indebted to Vittorio Colantuoni for generously providing the human RBP cDNA for this work. We thank Lila Gierasch for help with the stopped-flow fluorescence studies and use of the lab instrumentation. We are very grateful to Martin Flajnik, Mary Lou King, William Whelan, Walter Scott and the Brew Group for stimulating discussions regarding the relationship between conserved residues and folding. We acknowledge Jane Richardson for coining the term fold-determining core for our work. We acknowledge Yuji Goto for valuable discussion regarding the folding behavior of  $\beta$ -lactoglobulin. D.H. was supported by JSPS Postdoctoral Fellowships for Research Abroad. This work is in part a contribution from the Oxford Centre for Molecular Sciences which is supported by the BBSRC, EPSRC and MRC. This work was also funded by NIH Grants GM21363 to K.B. and GM27616 to L.M.G.

#### References

- [1] Levinthal, C. (1968) *J. Chim. Phys.* 65, 44–45.
- [2] Greene, L.H. and Brew, K. (1995) *Protein Eng.* 8 (Suppl.), 100.
- [3] Shakhnovich, E., Abkevich, V. and Pitsyn, O.B. (1996) *Nature* 379, 96–98.
- [4] Brew, K. and Greene, L.H. (1997) *Protein Eng.* 10 (Suppl.), 44.
- [5] Greene, L. H. (1998) Investigation into the relationship between sequence, structure and folding in a model lipocalin: human serum retinol-binding protein, Ph.D. Thesis, University of Miami, Miami, FL.
- [6] Mirny, L.A. and Shakhnovich, E.I. (1999) *J. Mol. Biol.* 291, 177–196.
- [7] Chothia, C. and Gerstein, M. (1997) *Nature* 335, 579–581.
- [8] Heidary, D.K. and Jennings, P.A. (2002) *J. Mol. Biol.* 316, 789–798.
- [9] Larson, S.M., Ruczinski, I., Davidson, A.R., Baker, D. and Plaxco, K.W. (2002) *J. Mol. Biol.* 316, 225–233.
- [10] Fowler, S.B. and Clarke, J. (2001) *Structure* 9, 355–366.
- [11] Gunasekaran, K., Eyles, S.J., Hagler, A.T. and Gierasch, L.M. (2001) *Curr. Opin. Struct. Biol.* 11, 83–93.
- [12] Nishimura, C., Prytulla, S., Dyson, J. and Wright, P.E. (2000) *Nat. Struct. Biol.* 7, 679–686.
- [13] Hamill, S.J., Steward, A. and Clarke, J. (2000) *J. Mol. Biol.* 297, 165–178.
- [14] Ortiz, A.R. and Skolnick, J. (2000) *Biophys. J.* 79, 1787–1799.
- [15] Kragelund, B.B., Osmark, P., Neergaard, T.B., Schiødt, J., Kristiansen, K., Knudsen, J. and Poulsen, F.M. (1999) *Nat. Struct. Biol.* 6, 594–601.
- [16] Parker, M.J., Dempsey, C.E., Hosszu, L.L.P., Waltho, J.P. and Clarke, A.R. (1998) *Nat. Struct. Biol.* 5, 194–198.
- [17] Flower, D.R. (1996) *Biochem. J.* 318, 1–14.
- [18] Pervaiz, S. and Brew, K. (1985) *Science* 228, 335–337.
- [19] Flower, D.R. (2000) *Biochim. Biophys. Acta* 1482, 46–56.
- [20] Bishop, R.E. and Weiner, J.H. (1996) *Trends Biochem. Sci.* 21, 127.

- [21] Kuwata, K., Shastry, R., Cheng, H., Hoshino, M., Batt, C.A., Goto, Y. and Roder, H. (2001) *Nat. Struct. Biol.* 8, 151–155.
- [22] Sander, C. and Schneider, R. (1991) *Proteins Struct. Funct. Genet.* 9, 56–68.
- [23] Levitt, M. (1976) *J. Mol. Biol.* 104, 59–107.
- [24] Greene, L.H., Chrysina, E.D., Irons, L.I., Papageorgiou, A., Acharya, K.R. and Brew, K. (2001) *Protein Sci.* 10, 2301–2316.
- [25] Bychkova, V.E., Berni, R., Rossi, G.L., Kutysenko, V.P. and Ptitsyn, O.B. (1992) *Biochemistry* 31, 7566–7571.
- [26] North, A.C.T. (1989) *Int. J. Biol. Macromol.* 6, 38–40.
- [27] Freifelder, D. (1982) in: *Physical Biochemistry – Applications to Biochemistry and Molecular Biology*, 2nd edn., pp. 537–572, W.H. Freeman and Co., New York.
- [28] Agashe, V.R., Shastry, M.C.R. and Udgaonkar, J.B. (1995) *Nature* 377, 754–757.
- [29] Woody, R.W. and Dunker, A.K. (1996) in: *Circular Dichroism and the Conformational Analysis of Biomolecules* (Fasman, G.D., Ed.), pp. 109–157, Plenum Press, New York.
- [30] Kuwajima, K. (1996) in: *Circular Dichroism and the Conformational Analysis of Biomolecules* (Fasman, G.D., Ed.), pp. 159–182, Plenum Press, New York.
- [31] Hamada, D., Segawa, S. and Goto, Y. (1996) *Nat. Struct. Biol.* 10, 868–873.
- [32] Arai, M., Ikura, T., Semisotnov, G.V., Kihara, H., Amemiya, Y. and Kuwajima, K. (1998) *J. Mol. Biol.* 275, 149–162.
- [33] Forge, V., Hoshino, M., Kuwata, K., Arai, M., Kuwajima, K., Batt, C.A. and Goto, Y. (2000) *J. Mol. Biol.* 296, 1039–1051.
- [34] Ragona, L., Catalano, M., Zetta, L., Longhi, R., Fogolari, F. and Molinari, H. (2002) *Biochemistry* 41, 2786–2796.
- [35] Hamada, D., Kuroda, Y., Tanaka, T. and Goto, Y. (1995) *J. Mol. Biol.* 254, 737–746.
- [36] Rost, B. (1996) *Methods Enzymol.* 266, 525–539.
- [37] Kneller, D.G., Cohen, F.E. and Langridge, R.L. (1990) *J. Mol. Biol.* 214, 171–182.
- [38] Combet, C., Blanchet, C., Geourjon, C. and Deleage, G. (2000) *Trends Biochem. Sci.* 291, 147–150.