Published online 17 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D429–D434***
doi: 10.1093/nar/gkv1185

Start2Fold: a database of hydrogen/deuterium exchange data on protein folding and stability

Rita Pancsa[1,2,*,†], Mihaly Varadi[1,2,*,†], Peter Tompa[1,2,3,4] and Wim F. Vranken[1,2,3,*]

[1]Structural Biology Brussels, Vrije Universiteit Brussel (VUB), Brussels 1050, Belgium, [2]Structural Biology Research Center (IB[2]), VIB, Brussels 1050, Belgium, [3]Interuniversity Institute of Bioinformatics in Brussels (IB[2]), ULB-VUB, Brussels 1050, Belgium and [4]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest 1113, Hungary

Received August 15, 2015; Revised October 2, 2015; Accepted October 24, 2015

ABSTRACT

Proteins fulfil a wide range of tasks in cells; understanding how they fold into complex three-dimensional (3D) structures and how these structures remain stable while retaining sufficient dynamics for functionality is essential for the interpretation of overall protein behaviour. Since the 1950's, solvent exchange-based methods have been the most powerful experimental means to obtain information on the folding and stability of proteins. Considerable expertise and care were required to obtain the resulting datasets, which, despite their importance and intrinsic value, have never been collected, curated and classified. Start2Fold is an openly accessible database (http://start2fold.eu) of carefully curated hydrogen/deuterium exchange (HDX) data extracted from the literature that is open for new submissions from the community. The database entries contain (i) information on the proteins investigated and the underlying experimental procedures and (ii) the classification of the residues based on their exchange protection levels, also allowing for the instant visualization of the relevant residue groups on the 3D structures of the corresponding proteins. By providing a clear hierarchical framework for the easy sharing, comparison and (re-)interpretation of HDX data, Start2Fold intends to promote a better understanding of how the protein sequence encodes folding and structure as well as the development of new computational methods predicting protein folding and stability.

INTRODUCTION

Proteins are dynamic molecules that display a wide range of behaviours and fulfil many functions in the cell; understanding how they fold into complex three-dimensional (3D) structures and how these structures remain stable is essential for the interpretation of their overall behaviour. However, studying protein folding and stability is a difficult task; the complexity of the folding process and the diversity of and interchange between possible protein conformations require highly time-sensitive and complex experiments. Only solvent exchange-based methods (1–6) and protein engineering approaches (7,8) are able to provide experimental information at the level of individual amino acid residues. Among these approaches, hydrogen/deuterium exchange (HDX) techniques are the most widely used (1,6); they are based on the notion that, for a given residue, the rate of proton exchange between its backbone amide proton and the solvent (water) provides information on the structural state of that residue (9). Amide protons of residues that form stable hydrogen bonds, and are often deeply buried in the protein, are protected from solvent and basically stop exchanging, while the amide protons of solvent exposed residues not involved in hydrogen bonds display little or no protection from exchange. These HDX techniques (i) can be applied under various conditions, (ii) have the potential to provide information on a large fraction of the protein residues while highlighting residue-specific details, (iii) are exquisitely sensitive to the dynamics of structure and rare conformational changes and (iv) in combination with rapid mixing techniques can be used to monitor folding events with good time resolution (10,11). Due to their indisputable advantages, over the years these techniques have developed into powerful tools for studying protein structure, stability and dynamics (12), through examining the native state (13,14), partially folded equilibrium intermediates (15–17), kinetic folding intermediates (18–20) and association reactions of proteins (21–23).

*To whom correspondence should be addressed. Tel: +32 2 6505943; Fax: +32 2 6291963; Email: wvranken@vub.ac.be
Correspondence may also be addressed to Rita Pancsa. Tel: +44 1223 267823; Fax: +44 1223 268300; Email: rpancsa@mrc-lmb.cam.ac.uk
Correspondence may also be addressed to Mihaly Varadi. Tel: +44 1223 494278; Fax: +44 1223 494468; Email: mvaradi@ebi.ac.uk
†These authors contributed equally to the paper as first authors.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Native exchange experiments investigate proteins in their folded or partially folded states and report on the stability of hydrogen bonds that exist in these states (13,14,16). By varying the environmental conditions, like pH or denaturant concentrations, such measurements can also provide quantitative information on the protection levels of the individual amide protons in function of changes in the conformation of the protein. By comparing these values, one can distinguish more stable from less stable regions of the protein fold (13,14,16), and can moreover provide clues about folding kinetics under certain conditions (20,24). The protection levels or exchange rates of amide protons can also be followed from the completely unfolded state throughout the entire course of folding of the proteins by a range of methods, for example pulsed labelling (25–27), quenched flow (28–30) and competition-based HDX measurements (10,31) (coupled with either nuclear magnetic resonance (NMR) or mass spectrometry (MS) as detection techniques); they indicate which regions of the protein first form local structure, thus providing invaluable information on the folding mechanisms of proteins.

Over the past 40 years numerous solvent-exchange experiments have been carried out to gain insights into protein structure, stability, folding and dynamics (32). Despite the invaluable insights these experiments have provided and their availability since the 1950's (33), the resulting datasets have never been assembled and made available as a database, which makes it difficult to draw general conclusions and/or to develop computational prediction tools. In our opinion this lack of a public HDX database is due to the heterogeneity of both the relevant methods and the measures used to describe the protection rates of residues (protection factors, folding rate constants, burst phase amplitude, midpoint of folding etc.), which make the collection and comparison of different HDX data in the literature especially difficult and labour intensive.

We overcame these problems and present here Start2Fold (http://start2fold.eu), a comprehensive collection of carefully curated and classified residue-level folding and stability data derived from solvent exchange-based experiments, including native-state, pulsed labelling, quenched flow and competition-based HDX experiments and oxidative labelling measurements.

MATERIALS AND METHODS

Data collection and classification

The measurements were collected from literature and subsequently classified based on (i) whether they investigate protein folding or stability and (ii) whether they provide information on the residue or segment level. Due to the heterogeneity of data, the published protection levels and rates cannot be directly compared between proteins and it is not possible to apply generic quantitative thresholds of protection. To overcome this problem and enable comparisons on a qualitative level, the proteins' residues were therefore classified into groups based on their detected exchange protection levels. We mostly adopted the residue classification that either was proposed in the original publication describing the measurement or that was suggested by Li and Woodward in their 1999 comparative analysis (6). In the few cases where the authors provided good quality measures of residue folding rates without a classification, we determined the protection thresholds according to our best knowledge so that ∼5–15% of all the residues were classified as the earliest folding ones. If the folding of the same protein was investigated by multiple measurements (e.g. horse cytochrome c (10,34,35)), the experiment with the best time resolution was selected, or the results of multiple measurements were retained for presentation in the database.

Database structure and content

Start2Fold is a relational database implemented in MySQL (http://www.mysql.com/). Data in Start2Fold is structured hierarchically in the following manner (Figure 1): on the top level (level 1) is the entry, which might be associated with one or multiple molecular systems (i.e. a protein of multiple chains or even complexes of multiple proteins). Each entry has a distinct identifier, which consists of a tag (STF) followed by a four letter code (e.g. STF0001). An entry might have several associated protein chains (level 2). This level (or class) stores information on the UniProt (36) and Protein Data Bank (PDB) (37,38) IDs of the protein chain, and serves to provide direct cross-links to these online repositories. Each protein chain might have several corresponding residue sets defined by their protection levels, which can range between early, intermediate and late for folding and strong, medium and weak for stability measurements (level 3). Additional information recorded on the residue sets include the resolution (residue-level or segment-level), the number of probes, the protection threshold, the experimental conditions (pH, temperature), the actual protein sequence used in the experiment (with any modifications/mutations), the PubMed ID (http://www.ncbi.nlm.nih.gov/pubmed) of the original publication and a textual description of the experimental procedure taken from the original publication. Finally, for the residues belonging to each residue set, the sequence position and amino acid types are also provided (level 4).

RESULTS

Currently, Start2Fold contains 57 entries with 219 residue sets defined based on the protection levels. These residue sets contain a total of 4172 residues, with the same residues appearing in multiple sets, since there are data available from both stability and folding measurements for most of the protein entries. Besides the measurement and residue classifications, all entries of the database contain structural features of the investigated protein, the detailed descriptions of the underlying experimental procedures, sample components, measurement conditions, cross references to other databases and references to the original publications. The accession pages of the entries also allow for the visualization of the relevant residue groups in the 3D structures of the corresponding proteins (where available). We also provide an XML template (39) that enables scientists to deposit new data into Start2Fold.

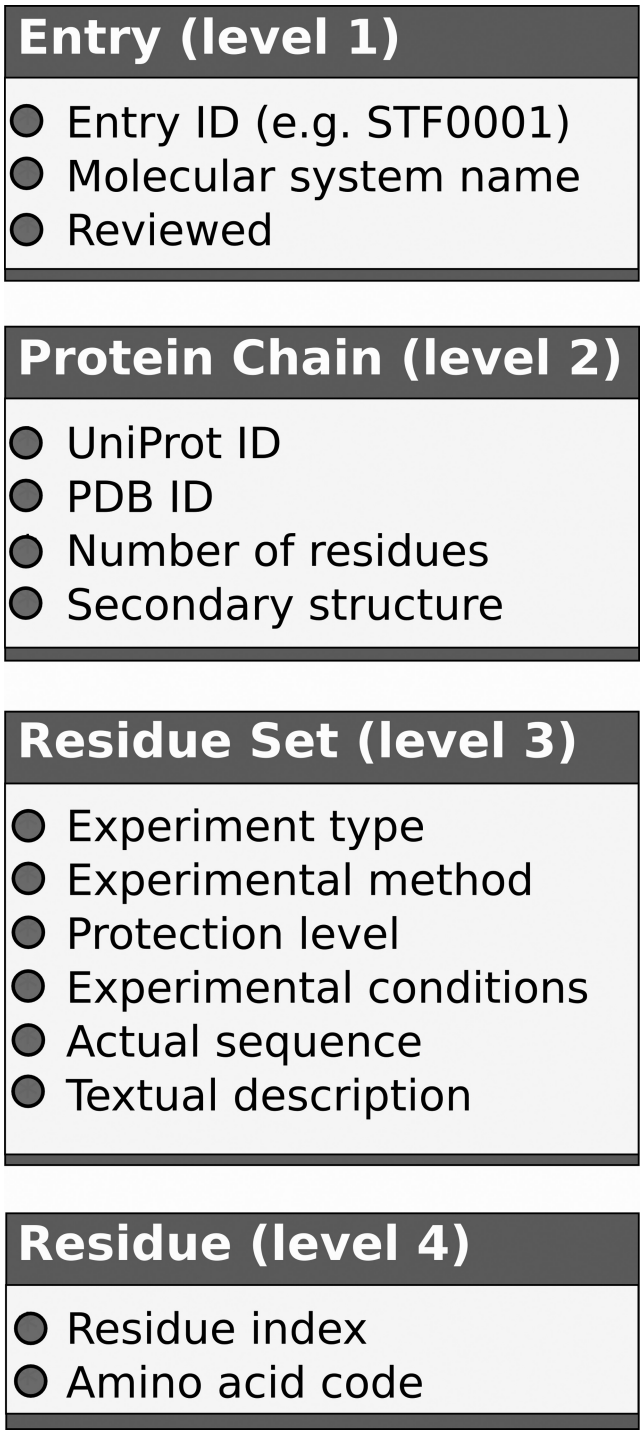

Figure 1. Database structure: Start2Fold is organized hierarchically into four levels: the entries (level 1) are on the top; each entry might have multiple associated protein chains (level 2); which in turn might have several experimental residue sets (level 3), with information on their associated residues (level 4). Each level records relevant information that is displayed on the accession screens or in the XML files of the entries.

User interface and website features

The user interface of Start2Fold is divided into five main sections. The 'home' section briefly introduces the database and the types of data contained within. It also provides the user with a contact form to send inquiries and feedback to the developers. Three sections provide browsing options by different criteria, i.e. proteins, residue sets or entries. Each option provides an ordered list of the entries that can be rearranged by the most relevant information depending on the browsing option. When browsing by entries, the entry ID, and molecular system name are displayed. In case of browsing by proteins, the UniProt name of the protein, the entry ID, the UniProt and PDB IDs, the length of the protein chain, the corresponding UniProt fragment, and the secondary structure type of the chain constitute the browsing list. Lastly, when browsing by residue sets, the protection level, molecular system name, entry ID, experiment type and method, and PubMed reference are displayed in the list. The user is forwarded to the accession screen by clicking on the entry ID links on either the browsing lists or on the search results list. This page provides all the relevant information associated with the entry, along with a static picture of the protein structure (ifavailable) which links to an integrated JSmol applet page (http://sourceforge.net/projects/jsmol/). Finally, the 'help' section contains the detailed documentation of the database, with an in-depth user guide describing all the functionalities of Start2Fold.

The 'home', 'help' and browsing pages can be accessed from all the pages using the menu on the top left section of the screen. Additionally, the database can be searched using the 'search' field located on the right side of the menu. Searching Start2Fold can be performed by typing in protein names, UniProt/PDB IDs, experiment types, experiment methods and protection levels in the search field, and pressing 'search'.

Accession screens

The actual information stored within Start2Fold is displayed on the accession screens (Figure 2). The entry ID and the title of the entry appear on the top of the page. Below is the 'download entry in xml' link, which provides the complete entry in XML format for downloading. This XML follows the structure of the XML template found on the welcome page. Alternatively, this XML can be directly accessed by adding '.xml' to the entry URL (e.g. http://start2fold.eu/STF0004.xml).

The integrated JSmol applet is available by either clicking on the image of the protein structure or by clicking on the "click here" link under the "visualize the data" section. This applet can be used to visualize the different residue sets or segments by clicking on one of the buttons (Figure 2A). The 'reset view' button can be clicked to reset the JSmol applet. Please note that due to technical issues with the current version of JSmol the loading speed can be slow on some browsers and depends on the available internet connection.

The protein information and experimental set sections are next to and below the visualization applet link (Figure 2B). The protein information tab provides the name of the protein, the species of origin, the number of residues in the protein chain and cross-links to UniProt and PDB. The experimental sets can be opened and closed by clicking on the 'show' and 'hide' buttons. These sections display the corresponding reference details, the experimental type (stability/folding) and method, the experimental conditions

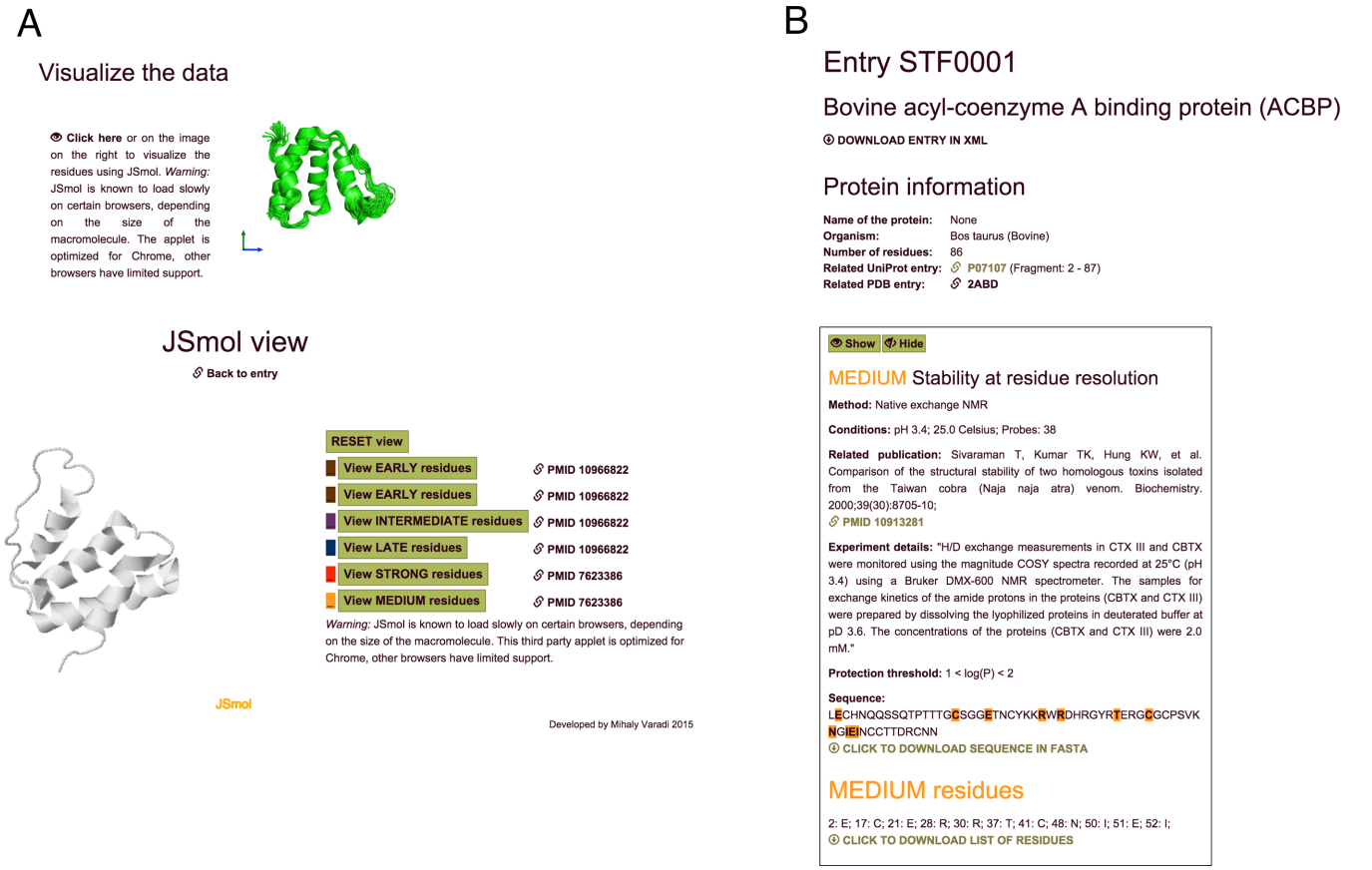

Figure 2. The accession screen: every piece of information that is stored in Start2Fold is displayed on the accession screens. The top section of these screens hosts an integrated JSmol applet for interactive visualization (panel **A**). General information of the protein chain(s) and relevant information on the experimental sets are displayed below (panel **B**). The complete entry can be downloaded in XML format, while the sequences can be retrieved in FASTA format and the residue/segment-specific information in the form of a list file by clicking on the respective links.

(pH, temperature, number of probes), a brief description of the experiment and the actual sequence that was used for the measurement. This sequence can be downloaded in FASTA format by clicking on the 'click to download sequence in fasta' link under the sequence. Alternatively, the sequences can be directly accessed by adding '.fasta' to the URL (e.g. http://start2fold.eu/STF0008.fasta). Finally, the residues are listed by their indices and their one-letter amino acid codes. Each residue of in a set is highlighted on the sequence to allow quick visualization at a glance. The residue lists can be downloaded either by clicking the 'click to download list of residues' link below the residues or by directly accessing them via adding '.residues' to the URL (e.g. http://start2fold.eu/STF0008.residues).

DISCUSSION

We hope that the collection, curation and integration of HDX data into a single well-organized searchable database, Start2Fold, will stimulate future structure/folding-related work, both at the experimental and computational level. It is likely that the heterogeneous nature of HDX data stalled the development of such a database, an issue which we here addressed by introducing a clear classification scheme of the residues based on their protection levels. Although proteins (un-)fold at varying overall rates, it is now possible to compare the order in which residues become (de-)protected, which will hopefully allow refinement of the relationship between the stability and folding cores of proteins and re-interpretation of accumulated HDX data, so providing valuable new insights. We encourage researchers to submit their new exchange data, and will consider including additional types of data into Start2Fold, such as high quality computational simulations of protein folding (40–43).

In all, we hope that Start2Fold will enable answering some of the remaining open questions related to the folding mechanisms of proteins (44–48), and that it will contribute to the development of new methods that allow for the reliable calculation of protein folding, structure and stability from sequence.

AVAILABILITY

Start2Fold is openly available for the scientific community at http://start2fold.eu. The database is open to submissions; we encourage users to submit their folding and stability data using the template XML provided on the welcome page of the website.

ACKNOWLEDGEMENT

We thank Palma Pakai for her practical advice on the graphical design of the online user interface.

FUNDING

Brussels Institute for Research and Innovation (Innoviris) [BB2B 2010-1-12 to W.F.V.]; Research Foundation—Flanders (FWO) [Odysseus grant number G.0029.12 to P.T.]. Funding for open access charge: Brussels Institute for Research and Innovation (Innoviris) [BB2B 2010-1-12 to W.F.V.]; Research Foundation—Flanders (FWO) [Odysseus grant number G.0029.12 to P.T.].
Conflict of interest statement. None declared.

REFERENCES

1. Bai,Y. (2006) Protein folding pathways studied by pulsed- and native-state hydrogen exchange. *Chem. Rev.*, **106**, 1757–1768.
2. Jackson,S.E. (1998) How do small single-domain proteins fold? *Fold Des.*, **3**, R81–R91.
3. Konermann,L., Pan,J. and Liu,Y.H. (2011) Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.*, **40**, 1224–1234.
4. Konermann,L., Pan,Y. and Stocks,B.B. (2011) Protein folding mechanisms studied by pulsed oxidative labeling and mass spectrometry. *Curr. Opin. Struct. Biol.*, **21**, 634–640.
5. Konermann,L., Stocks,B.B., Pan,Y. and Tong,X. (2010) Mass spectrometry combined with oxidative labeling for exploring protein structure and folding. *Mass Spectrom. Rev.*, **29**, 651–667.
6. Li,R. and Woodward,C. (1999) The hydrogen exchange core and protein folding. *Protein Sci.*, **8**, 1571–1590.
7. Fersht,A.R., Matouschek,A. and Serrano,L. (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.*, **224**, 771–782.
8. Naganathan,A.N. and Munoz,V. (2010) Insights into protein folding mechanisms from large scale analysis of mutational effects. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 8611–8616.
9. Hvidt,A. and Nielsen,S.O. (1966) Hydrogen exchange in proteins. *Adv. Protein Chem.*, **21**, 287–386.
10. Fazelinia,H., Xu,M., Cheng,H. and Roder,H. (2014) Ultrafast hydrogen exchange reveals specific structural events during the initial stages of folding of cytochrome c. *J. Am. Chem. Soc.*, **136**, 733–740.
11. Roder,H. and Wuthrich,K. (1986) Protein folding kinetics by combined use of rapid mixing techniques and NMR observation of individual amide protons. *Proteins*, **1**, 34–42.
12. Krishna,M.M., Hoang,L., Lin,Y. and Englander,S.W. (2004) Hydrogen exchange methods to study protein folding. *Methods*, **34**, 51–64.
13. Kragelund,B.B., Knudsen,J. and Poulsen,F.M. (1995) Local perturbations by ligand binding of hydrogen deuterium exchange kinetics in a four-helix bundle protein, acyl coenzyme A binding protein (ACBP). *J. Mol. Biol.*, **250**, 695–706.
14. Merstorf,C., Maciejak,O., Mathe,J., Pastoriza-Gallego,M., Thiebot,B., Clement,M.J., Pelta,J., Auvray,L., Curmi,P.A. and Savarin,P. (2012) Mapping the conformational stability of maltose binding protein at the residue scale using nuclear magnetic resonance hydrogen exchange experiments. *Biochemistry*, **51**, 8919–8930.
15. Eliezer,D., Jennings,P.A., Dyson,H.J. and Wright,P.E. (1997) Populating the equilibrium molten globule state of apomyoglobin under conditions suitable for structural characterization by NMR. *FEBS Lett.*, **417**, 92–96.
16. Hughson,F.M., Wright,P.E. and Baldwin,R.L. (1990) Structural characterization of a partly folded apomyoglobin intermediate. *Science*, **249**, 1544–1548.
17. Nishimura,C., Dyson,H.J. and Wright,P.E. (2008) The kinetic and equilibrium molten globule intermediates of apoleghemoglobin differ in structure. *J Mol Biol*, **378**, 715–725.
18. Kern,G., Handel,T. and Marqusee,S. (1998) Characterization of a folding intermediate from HIV-1 ribonuclease H. *Protein Sci.*, **7**, 2164–2174.
19. Pan,J., Han,J., Borchers,C.H. and Konermann,L. (2010) Characterizing short-lived protein folding intermediates by top-down hydrogen exchange mass spectrometry. *Anal. Chem.*, **82**, 8591–8597.
20. Parker,M.J. and Marqusee,S. (2001) A kinetic folding intermediate probed by native state hydrogen exchange. *J. Mol. Biol.*, **305**, 593–602.
21. Mobley,J.A. and Poliakov,A. (2009) Detection of early unfolding events in a dimeric protein by amide proton exchange and native electrospray mass spectrometry. *Protein Sci.*, **18**, 1620–1627.
22. Pan,J., Rintala-Dempsey,A.C., Li,Y., Shaw,G.S. and Konermann,L. (2006) Folding kinetics of the S100A11 protein dimer studied by time-resolved electrospray mass spectrometry and pulsed hydrogen-deuterium exchange. *Biochemistry*, **45**, 3005–3013.
23. Simler,B.R., Levy,Y., Onuchic,J.N. and Matthews,C.R. (2006) The folding energy landscape of the dimerization domain of Escherichia coli Trp repressor: a joint experimental and theoretical investigation. *J. Mol. Biol.*, **363**, 262–278.
24. Arrington,C.B. and Robertson,A.D. (1997) Microsecond protein folding kinetics from native-state hydrogen exchange. *Biochemistry*, **36**, 8686–8691.
25. Udgaonkar,J.B. and Baldwin,R.L. (1990) Early folding intermediate of ribonuclease A. *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 8197–8201.
26. Uzawa,T., Nishimura,C., Akiyama,S., Ishimori,K., Takahashi,S., Dyson,H.J. and Wright,P.E. (2008) Hierarchical folding mechanism of apomyoglobin revealed by ultra-fast H/D exchange coupled with 2D NMR. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 13859–13864.
27. Walkenhorst,W.F., Edwards,J.A., Markley,J.L. and Roder,H. (2002) Early formation of a beta hairpin during folding of staphylococcal nuclease H124L as detected by pulsed hydrogen exchange. *Protein Sci.*, **11**, 82–91.
28. Greene,L.H., Li,H., Zhong,J., Zhao,G. and Wilson,K. (2012) Folding of an all-helical Greek-key protein monitored by quenched-flow hydrogen-deuterium exchange and NMR spectroscopy. *Eur. Biophys. J.*, **41**, 41–51.
29. Hsieh,H.C., Kumar,T.K., Sivaraman,T. and Yu,C. (2006) Refolding of a small all beta-sheet protein proceeds with accumulation of kinetic intermediates. *Arch. Biochem. Biophys.*, **447**, 147–154.
30. Varley,P., Gronenborn,A.M., Christensen,H., Wingfield,P.T., Pain,R.H. and Clore,G.M. (1993) Kinetics of folding of the all-beta sheet protein interleukin-1 beta. *Science*, **260**, 1110–1113.
31. Kato,H., Vu,N.D., Feng,H., Zhou,Z. and Bai,Y. (2007) The folding pathway of T4 lysozyme: an on-pathway hidden folding intermediate. *J. Mol. Biol.*, **365**, 881–891.
32. Englander,S.W. and Kallenbach,N.R. (1983) Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Q. Rev. Biophys.*, **16**, 521–655.
33. Englander,S.W., Mayne,L., Bai,Y. and Sosnick,T.R. (1997) Hydrogen exchange: the modern legacy of Linderstrom-Lang. *Protein Sci.*, **6**, 1101–1109.
34. Roder,H., Elove,G.A. and Englander,S.W. (1988) Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR. *Nature*, **335**, 700–704.
35. Yang,H. and Smith,D.L. (1997) Kinetics of cytochrome c folding examined by hydrogen exchange and mass spectrometry. *Biochemistry*, **36**, 14992–14999.
36. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
37. Berman,H.M., Kleywegt,G.J., Nakamura,H. and Markley,J.L. (2014) The Protein Data Bank archive as an open data resource. *J. Comput. Aided Mol. Des.*, **28**, 1009–1014.
38. Gutmanas,A., Alhroub,Y., Battle,G.M., Berrisford,J.M., Bochet,E., Conroy,M.J., Dana,J.M., Fernandez Montecelo,M.A., van Ginkel,G., Gore,S.P. *et al.* (2014) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **42**, D285–D291.
39. Jackson,K. (2000) XML: Extensible markup language. *Tech. Commun.*, **47**, 108–109.
40. Compiani,M. and Capriotti,E. (2013) Computational and theoretical methods for protein folding. *Biochemistry*, **52**, 8601–8624.
41. Lane,T.J., Shukla,D., Beauchamp,K.A. and Pande,V.S. (2013) To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.*, **23**, 58–65.
42. Lindorff-Larsen,K., Piana,S., Dror,R.O. and Shaw,D.E. (2011) How fast-folding proteins fold. *Science*, **334**, 517–520.

43. Voelz,V.A., Bowman,G.R., Beauchamp,K. and Pande,V.S. (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1–39). *J. Am. Chem. Soc.*, **132**, 1526–1528.
44. Daggett,V. and Fersht,A.R. (2003) Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.*, **28**, 18–25.
45. Dill,K.A. and MacCallum,J.L. (2012) The protein-folding problem, 50 years on. *Science*, **338**, 1042–1046.
46. Englander,S.W. and Mayne,L. (2014) The nature of protein folding pathways. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 15873–15880.
47. Haran,G. (2012) How, when and why proteins collapse: the relation to folding. *Curr. Opin. Struct. Biol.*, **22**, 14–20.
48. Sosnick,T.R. and Barrick,D. (2011) The folding of single domain proteins–have we reached a consensus? *Curr. Opin. Struct. Biol.*, **21**, 12–24.