Centenary Award and Sir Frederick Gowland Hopkins Memorial Lecture

Protein folding, structure prediction and design

David Baker[*1]

[*]Department of Biochemistry, University of Washington/HMMI, Seattle, WA 98195, U.S.A.

Centenary Award and Frederick Gowland Hopkins Memorial Lecture
Delivered at the MRC Laboratory of Molecular Biology, Cambridge, on 13 December 2012
David Baker

Abstract

I describe how experimental studies of protein folding have led to advances in protein structure prediction and protein design. I describe the finding that protein sequences are not optimized for rapid folding, the contact order–protein folding rate correlation, the incorporation of experimental insights into protein folding into the Rosetta protein structure production methodology and the use of this methodology to determine structures from sparse experimental data. I then describe the inverse problem (protein design) and give an overview of recent work on designing proteins with new structures and functions. I also describe the contributions of the general public to these efforts through the Rosetta@home distributed computing project and the FoldIt interactive protein folding and design game.

I was reminded by the citation for the Centenary Award of how much my research group's interests have changed since I started at the University of Washington in 1994. In the present article, I describe how this occurred. A recurring theme is that, in research, one should not plan too far ahead, as the most interesting discoveries tend to be the most unexpected.

Part of my postdoctoral work had been studying a protein which, unlike its structurally related cousins, folded extremely slowly, with a half-time of 1 year or more [1], and I imagined that, for any given protein fold, there were likely to exist sequences with folding rates that spanned a very broad range. To understand how sequences determined folding rates and mechanisms, it seemed logical to start with the simplest possible cases of protein folding; since the combinatorial complexity of folding increases exponentially with chain length, this meant focusing on the smallest autonomously folding protein domains. Starting out at the University of Washington, I chose two small (~60 residues) proteins as model systems, and set out to obtain widely divergent sequences which folded up to these structures. We developed a phage display selection system which allowed selection of sequences which retained the ability to fold to these structures from very large randomized libraries [2]. In the case of the SH3 (Src homology 3) domain, some of the new sequences that retained the ability to fold were constituted almost entirely from a five-letter amino acid alphabet [3].

With these widely divergent sequence libraries in hand, we were set to investigate the extent to which sequence determines protein folding rates. There had been considerable discussion of how protein-folding pathways and mechanisms had been encoded in protein sequences by natural selection, and, if this were the case, the heavily mutated sequences would be expected to fold more slowly than their naturally occurring optimized counterparts. What we found, however, was quite the opposite. Whereas the selected random sequences were almost always less stable than the naturally occurring ones, their folding rates were as often higher as lower [4]. This showed clearly that amino acid sequences are not optimized for rapid folding.

The plan of gaining insight into protein folding by studying very-slow-folding variants was clearly not going to fly. Instead, since protein-folding rates were evidently not determined by the details of the amino acid sequence,

Key words: biophysics, energy landscape, FoldIt, protein folding, Rosetta@home, structure determination.

[1]email dabaker@uw.edu

we considered possible alternatives. In a simple model where protein folding is a trade-off between the formation of attractive native interactions and the loss of chain configurational entropy, the determinant of the height of the free energy barrier to folding is the extent to which the formation of attractive interactions early in folding can compensate for the entropy loss. Attractive interactions between residues nearby in the amino acid sequence can be formed without greatly restricting the number of conformations available to the polypeptide chain, whereas formation of favourable interactions between residues distant along the sequence considerably reduces the possible configurations of the intervening chain segment. Hence we reasoned that proteins with interactions primarily between residues close along the sequence might fold more rapidly than proteins with interactions primarily between residues distant along the sequence. For a large set of proteins of known structure whose folding rates had been determined, we computed the average sequence separation between residues in contact in the structure (the contact order) and found that there was indeed a strong correlation between folding rate and the sequence separation between contacting residues with low-contact-order proteins folding orders of magnitude faster than high-contact-order proteins [5].

In studying the folding of the small model proteins, we had made several other observations that shaped what we did next. First, there were segments of local structure that were stable as isolated peptides, but most peptide sequences derived from protein sequences had little persistent structure [6]. Secondly, mutations in certain turn regions and in the protein core lowered the folding rate [7,8]. These and other results suggested a picture of protein folding in which each segment of the polypeptide chain sampled a range of local conformations consistent with its local amino acid sequence, and folding occurred when these segments sampled the correct structure and orientation so as to bury the non-polar residues in a hydrophobic core.

I had been interested in the *ab initio* protein structure prediction problem since I had first learned about it, and we set out to implement what we had learned about protein folding in a structure-prediction method. The key assumption was that the ensemble of local structures sampled by a sequence segment during folding could be approximated by the ensemble of local structures that the sequence segment adopted in known protein structures. We searched through the conformational space defined by combinations of local structure possibilities using a simple Monte Carlo sampling protocol guided by an energy function capturing hydrophobic burial and backbone hydrogen-bonding. To make the calculations tractable, we used a simplified model in which each side chain was represented by a single sphere. We found that this approach, which graduate student Kim Simons called Rosetta, could rapidly fold small proteins up into compact three-dimensional structures with hydrophobic cores and that these structures, in some cases, were quite close to the experimentally observed structures [9].

We found that different folding trajectories with Rosetta ended up in different conformational minima, and hence, given a sequence, we carried out many independent trajectories and clustered the resulting structures to identify the broadest minima [10]. Although the largest cluster was often close to the native structure, in some cases one of the other clusters was a better model. Studies of the effects of point mutations on protein stability had shown that, whereas folding kinetics did not depend on the details of the sequence, protein stability certainly did, and to allow more accurate modelling and to distinguish between the alternative minima, we extended Rosetta to add on all the side-chain atoms and then minimize the energy with respect to all side-chain and backbone degrees of freedom simultaneously.

When experimentally determined structures and the models generated with Rosetta folding trajectories were refined using this all-atom model, we found that the native structure was almost always lower in energy than alternative topologies found by Rosetta. However, we soon found that this decrease in energy only occurred within 2–3 Å (1 Å = 0.1 nm) RMSD of the native structure, and only very rarely did Rosetta *ab initio* folding trajectories get this close. To enable the more comprehensive searching necessary to find the low-energy native energy minimum, we decided to enlist the help of the general public. We created a distributed computing project called Rosetta@home (available from http://boinc.bakerlab.org) in which volunteers donate spare cycles on their computers to carry out folding trajectories. Since this time, Rosetta@home volunteers have made absolutely invaluable contributions to our research projects; there are now on average 40000 computer processors active in this work which exceeds by far our local computing resources.

Following the development of the Rosetta structure-prediction methodology for monomeric proteins, we applied a similar approach to protein–protein docking [11], membrane protein structure prediction [12], symmetrical oligomer assembly [13] and RNA folding [14]. In all cases, we found, as in the case of monomeric soluble proteins, that the native structure was at a pronounced energy minimum compared with non-native structures generated with Rosetta, and that structure prediction was possible if we could sample close enough to the native structure to fall into this minimum. It is likely that this universal behaviour reflects a fundamental feature of the free energy landscapes of biological macromolecules which gives rise to their remarkable ability to self-organize. Since the number of non-native states accessible to a polypeptide chain, for example, is vast, there is a huge entropy cost in folding. To overcome this entropy cost, the energy of the native structure must be very much lower than the non-native structures. The ubiquitous native energy gap observed for every case of macromolecular self organization that we studied suggests that the magnitude of the actual gap is significantly larger than the errors in our energy calculations (which could still be quite substantial) [15].

Even with Rosetta@home, only in a small subset of cases could we sample closely enough to the native structure for any of the above biomolecular systems to accurately predict structure. Hence, unfortunately, *ab initio* prediction of macromolecular structure was (and still is) not a reliable way to determine macromolecular structures. However, we found that the structure modelling/prediction methodology we had developed suddenly became useful when combined with sparse experimental data to guide the search. This use of experimental data is very different from that in standard structure-determination methods: the experimental data need only point to the location of the global minimum, rather than completely specify the positions of the atoms in the structure. The utility of even very sparse amounts of data when searching a large space is illustrated by the problem of finding the lowest elevation point on Earth, the single piece of information that it does not lie in North America would eliminate lots of time wasted around Death Valley, and the additional piece of information that it is in the Middle East would greatly speed locating the Dead Sea. Rosetta is now being used routinely to solve structures with limited NMR (CS-Rosetta) [16] and X-ray diffraction data (MR-Rosetta) [17].

With protein all-atom modelling in place, we could suddenly approach a completely new class of problems. Rather than search for the lowest-energy structure for a given amino acid sequence, one could search for the lowest-energy sequence for a given structure. This is the protein design problem: given a structure or function of interest, design an amino acid sequence which folds to the structure/has the desired function. Brian Kuhlman developed efficient algorithms for finding the lowest-energy sequence for a given structure, and we were off and running. After redesigning a number of naturally occurring proteins, Brian took a big step forward and designed a protein with a new topology: TOP7 [18].

Once the design of monomeric proteins was established, the same methods could be applied to redesign protein–protein interaction affinity and specificity. We found that, using protein design calculations, it was possible to create orthogonal pairs of interacting proteins [19], and to create new proteins by designing domain–domain interfaces [20].

One of the most amazing things proteins do is to catalyse chemical reactions with very high efficiencies. We sought to create new catalysts by *de novo* computational design. Our approach involved, first, designing an ideal active site consisting of the transition state for the chemical reaction surrounded by disembodied protein functional groups in orientations optimal for catalysis, and, secondly, the design of proteins containing these sites. Using this approach, we were able to create catalysts for five different chemical reactions [21,22]. The activities of the designed enzymes were pretty low, but they could be improved considerably by directed evolution.

Protein design is particularly well suited for problems that Nature never tried to solve during natural evolution. Vaccine design is attractive for this reason, as no protein in Nature was under selective pressure to be an optimal antigen for eliciting a strong and specific immune response; indeed, most pathogen surface proteins are subjected to exactly the opposite pressure. Given a crystal structure of an epitope of a pathogenic protein, we approached vaccine design by designing proteins which stabilize the epitope in the conformation bound by the antibody. The designed proteins bind to antibodies which neutralize HIV, and their potential for eliciting these or similar antibodies when used as vaccines is currently being investigated.

A grand challenge is creating new binding proteins *de novo* for use in therapeutics and diagnostics. We developed general methods for designing proteins which bind with high affinity and specificity to sites of interest on proteins of known structure [23]. We used these methods to design small proteins which bind very tightly to the influenza virus haemagglutinin which is exposed on the outer surface of the virus (Figure 1). The designed proteins prevent the influenza virus from infecting cells in culture, and hence they are potential anti-flu therapeutics [24]. The computational methods are completely general, and we are currently designing small proteins to bind to sites of therapeutic importance on a number of protein targets (both human and pathogen). We are excited about the possibility that this may provide a general route to a new class of protein therapeutics intermediate in size between small-molecule and antibody drugs. The University of Washington has started up an Institute for Protein Design to investigate this possibility more vigorously.

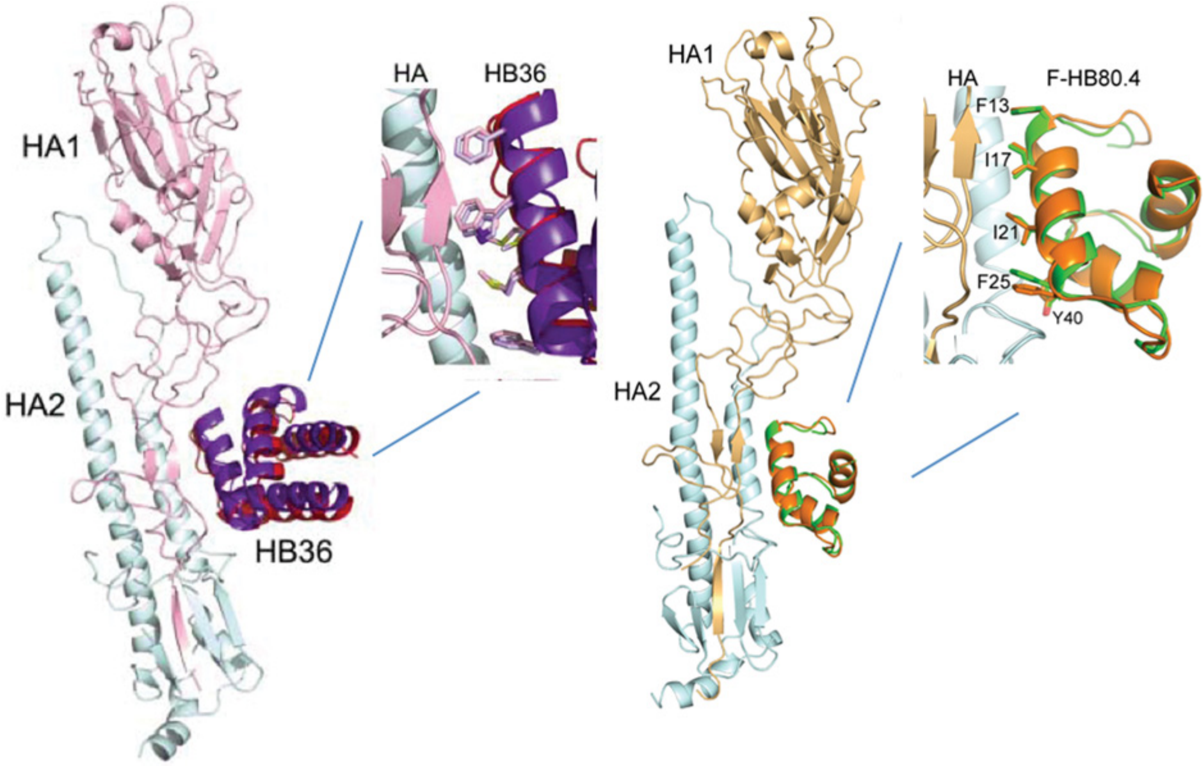
Self-assembling protein materials carry out a wide variety of functions in Nature: from viral capsids to cytoskeleton to silk. If we could engineer self-assembling protein-based materials to order, there would be many possible applications. Neil King developed an approach to designing regular polyhedra that utilizes building blocks that have cyclic symmetries found in the assembly. For example, to build a cubical octahedral structure, which has three-fold symmetry axes at the eight corners of the cube, we place a trimeric building block at each corner, dock them together, respecting the octahedral symmetry, and design the resulting interfaces between the trimers to stabilize the octahedral structure. We used this approach to design tetrahedral and octahedral structures, and are currently developing approaches to build more complex materials and exploring applications in vaccine design and drug delivery [25].

As the building blocks for new materials, we would ultimately like to use building blocks crafted *de novo* for this purpose. Natural proteins almost always have non-ideal features owing to selection for function. Nobu and Rie Koga identified general principles that allow the design of very stable proteins made of β-sheets and α-helices with very high accuracy, and used it to design a number of very stable brand new structures [26]; we are currently exploring a variety of ways of combining them into larger structures.

Rosetta@home volunteers led us into a completely new area a few years ago. When you run Rosetta@home on your computer, as of course you should, a screensaver pops up that shows the course of the calculation being done (a protein

Figure 1 | Computational design of influenza-binding proteins

Shown are two examples of small proteins (HB36 and F-HB80.4) designed to bind with high affinity to the conserved stem region of the influenza protein haemagglutinin (HA). In each example, the protein crystal structure (red, orange) is superimposed on the computational design (purple, green). Close-ups of the interfaces highlight the close agreement between the design and the crystal structure.



folding up, an interface being designed or whatever problem we are working on in the laboratory and sending out to the public for help with). Several volunteers wrote in a few years ago saying that, after watching Rosetta fold proteins up on their screensavers, they thought that it was in some cases inefficient and they could do better if there was some way for them to guide the protein as it folded. To enable this, we teamed up with the University of Washington computer science department, and developed an online multiplayer computer game called FoldIt which provides an interactive game interface to the Rosetta optimization algorithms and energy function [27]. In FoldIt, players compete to find the lowest-energy (highest score) solution to protein-structure prediction and design problems that we pose. FoldIt players in the last 2 years have made a number of quite important contributions: they solved the structure of a retroviral protease [28], developed new algorithms for finding low-energy protein conformations [29] and improved a *de novo* designed enzyme by rather large-scale redesign of the active site [30].

With the improvements in design methodology in the last several years, we can now design proteins for an ever-expanding range of applications. I am very excited about exploring this whole new world of possibilities in the years ahead, most of all the ones I cannot currently imagine.

Acknowledgements

I have been incredibly fortunate to have had many fantastic co-workers over the years. I do not have space here to acknowledge everybody who contributed to the work described above, so the following is necessarily highly incomplete. David Kim and David Riddle showed that protein-folding rates were robust to large-scale sequence changes, Kevin Plaxco and Kim Simons found the contact order–folding rate correlation, Kim Simons started (and named) Rosetta, David Kim created Rosetta@home, and Phil Bradley, Rhiju Das, Oliver Lange and Frank DiMaio showed the ubiquity of the native energy gap and extended Rosetta to solve structures using sparse data. In the protein design area, Brian Kuhlman and Gautam Dantas, and more recently Nobu and Rie Koga created novel proteins, Tanja Kortemme created new protein–protein interfaces, Lin Jiang, Alex Zanghellini, Daniela Roethlisberger and Eric Althoff created new enzymes, Bill Schief and his group created new vaccine candidates, Sarel Fleishman and Tim Whitehead created new influenza inhibitors, and Neil King and Will Sheffler created new self-assembling protein materials. FoldIt is a collaboration with University of Washington Center for Game Science. And I again thank all of the wonderful scientists I have been so privileged to work with over the years.

References

1 Baker, D., Sohl, J.L. and Agard, D.A. (1992) A protein-folding reaction under kinetic control. Nature **356**, 263–265

2 Gu, H., Yi, Q., Bray, S.T., Riddle, D.S., Shiau, A.K. and Baker, D. (1995) A phage display system for studying the sequence determinants of protein folding. Protein Sci. **4**, 1108–1117

3 Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, Q. and Baker, D. (1997) Functional rapidly folding proteins from simplified amino acid sequences. Nat. Struct. Biol. **4**, 805–809

4 Kim, D.E., Gu, H. and Baker, D. (1998) The sequences of small proteins are not extensively optimized for rapid folding by natural selection. Proc. Natl. Acad. Sci. U.S.A. **95**, 4982–4986

5 Plaxco, K.W., Simons, K.T. and Baker, D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. **277**, 985–994

6 Yi, Q. and Baker, D. (1996) Direct evidence for a two-state protein unfolding transition from hydrogen–deuterium exchange, mass spectrometry, and NMR. Protein Sci. **5**, 1060–1066

7 Kim, D.E., Yi, Q., Gladwin, S.T., Goldberg, J.M. and Baker, D. (1998) The single helix in protein l is largely disrupted at the rate-limiting step in folding. J. Mol. Biol. **284**, 807–815

8 Grantcharova, V.P., Riddle, D.S., Santiago, J.V. and Baker, D. (1998) Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. Nat. Struct. Biol. **5**, 714–720

9 Simons, K.T., Kooperberg, C., Huang, E. and Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. **268**, 209–225

10 Shortle, D., Simons, K.T. and Baker, D. (1998) Clustering of low-energy conformations near the native structures of small proteins. Proc. Natl. Acad. Sci. U.S.A. **95**, 11158–11162

11 Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A. and Baker, D. (2003) Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J. Mol. Biol. **331**, 281–299

12 Barth, P., Schonbrun, J. and Baker, D. (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. Proc. Natl. Acad. Sci. U.S.A. **104**, 15682–15687

13 Andre, I., Bradley, P., Wang, C. and Baker, D. (2007) Prediction of the structure of symmetrical protein assemblies. Proc. Natl. Acad. Sci. U.S.A. **104**, 17656–17661

14 Das, R. and Baker, D. (2007) Automated *de novo* prediction of native-like RNA tertiary structures. Proc. Natl. Acad. Sci. U.S.A. **104**, 14664–14669

15 Fleishman, S.J. and Baker, D. (2012) Role of the biomolecular energy gap in protein design, structure, and evolution. Cell **149**, 262–273

16 Raman, S., Lange, O.F., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G., Ramelot, T., Eletsky, A., Szyperski, T. et al. (2010) NMR structure determination for larger proteins using backbone-only data. Science **327**, 1014–1018

17 DiMaio, F., Terwilliger, T.C., Read, R.J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fas, D., Axelrod, H.L. et al. (2011) Improved molecular replacement by density- and energy-guided protein structure optimization. Nature **473**, 540–543

18 Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L. and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. Science **302**, 1364–1368

19 Kortemme, T., Joachimiak, L.A., Bullock, A.N., Schuler, A.D., Stoddard, B.L. and Baker, D. (2004) Computational redesign of protein–protein interaction specificity. Nat. Struct. Mol. Biol. **11**, 371–379

20 Chevalier, B.S., Kortemme, T., Chadsey, M.S., Baker, D., Monnat, R.J. and Stoddard, B.L. (2002) Design, activity, and structure of a highly specific artificial endonuclease. Mol. Cell **10**, 895–905

21 Röthlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O. et al. (2008) Kemp elimination catalysts by computational enzyme design. Nature **453**, 190–195

22 Jiang, L., Althoff, E.A., Clemente, F.R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J.L., Betker, J.L., Tanaka, F., Barbas, 3rd, C.F. et al. (2008) *De novo* computational design of retro-aldol enzymes. Science **319**, 1387–1391

23 Fleishman, S.J., Whitehead, T.A., Ekiert, D.C., Dreyfus, C., Corn, J.E., Strauch, E.-M., Wilson, I.A. and Baker, D. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science **332**, 816–821

24 Whitehead, T.A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S.J., De Mattos, C., Myers, C.A., Kamisetty, H., Blair, P., Wilson, I.A. and Baker, D. (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat. Biotechnol. **30**, 543–548

25 King, N.P., Sheffler, W., Sawaya, M.R., Vollman, B.S., Sumida, J.P., Andre, I., Gonen, T., Yeates, T.O. and Baker, D. (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science **336**, 1171–1174

26 Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T.B., Montelione, G.T. and Baker, D. (2012) Principles for designing ideal protein structures. Nature **419**, 222–227

27 Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D. and Popović, Z. (2010) Predicting protein structures with a multiplayer online game. Nature **466**, 756–760

28 Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z. et al. (2011) Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nat. Struct. Mol. Biol. **18**, 1175–1177

29 Khatib, F., Cooper, S., Tyka, M.D., Xu, K., Makedon, I., Popović, Z. and Baker, D. (2011) Algorithm discovery by protein folding game players. Proc. Natl. Acad. Sci. U.S.A. **108**, 18949–18953

30 Eiben, C.B., Siegel, J.B., Bale, J.B., Cooper, S., Khatib, F., Shen, B.W., Stoddard, B.L., Popović, Z. and Baker, D. (2012) Increased Diels–Alderase activity through backbone remodeling guided by FoldIt players. Nat. Biotechnol. **30**, 190–192

Received 12 February 2014
doi:10.1042/BST20130055