Prediction of chemical shifts for protein (mis-)folding studies

UNIVERSITY OF CAMBRIDGE

Kai Kohlhoff, Department of Chemistry

1. Introduction

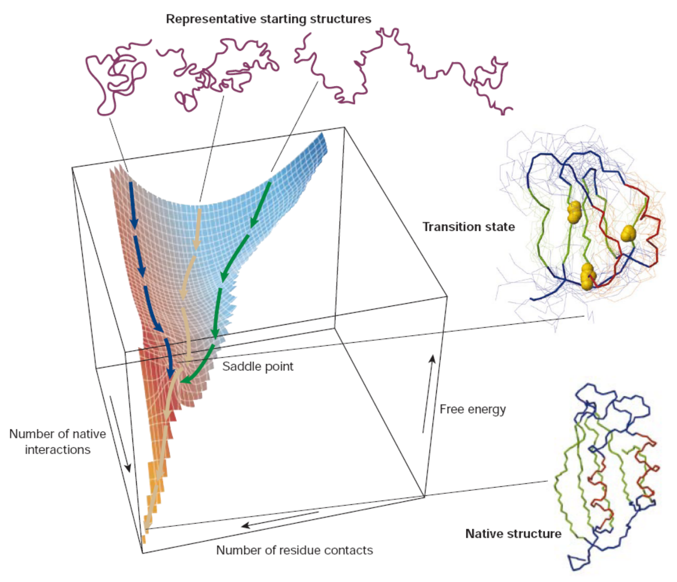

Fig. 1: Schematic representation of a simple energy landscape. Different starting structures of the same protein travel along the landscape towards their native structure, possibly via intermediates, or transition states. (from [2])

Proteins are amino acid chains that are formed at the ribosome. While initially unfolded, the chains usually have to fold into a specific three-dimensional 'native' structure in order to become biologically active [1]. This **folding** event, leading from a state of high energy to a minimum energy state, can be described as a pathway on a '**free energy landscape**' (figure 1): like a skier the protein tries to get from a mountain peak into the valley without having to put in additional energy. In some cases there might be several favourable pathways to do so. However, especially when with age the body's quality control mechanisms start to weaken, proteins might choose an incorrect folding pathway, **mis-fold**, and form **aggregates**: they end up in the wrong valley from where there is no return.

Such aggregates are believed to trigger diseases like **Alzheimer's** and **late onset diabetes** [3].

2. Research Project

Computational models can help defining energy landscapes and further our understanding of mis-folding events [4]. The behaviour of very small molecules can be treated via quantum mechanical calculations, but for more complex proteins it can only be approximated, for example via **Molecular Dynamics (MD)** computer simulations (figure 2). These methods rely on the laws of Physics and information obtained from experiments to approximate atomic movements.

We try to increase accuracy and speed of MD simulations by using experimental data from **Nuclear Magnetic Resonance spectroscopy (NMR)** [6].

In particular, we presently try predicting a property called 'chemical shifts' (c.s.) which is a very sensitive measure of an atom's atomic environment [7]. By defining penalty functions that depend on the difference between calculated c.s. of a simulated protein structure and experimental c.s of a target structure we hope to be able to guide computer-based folding towards that target structure.

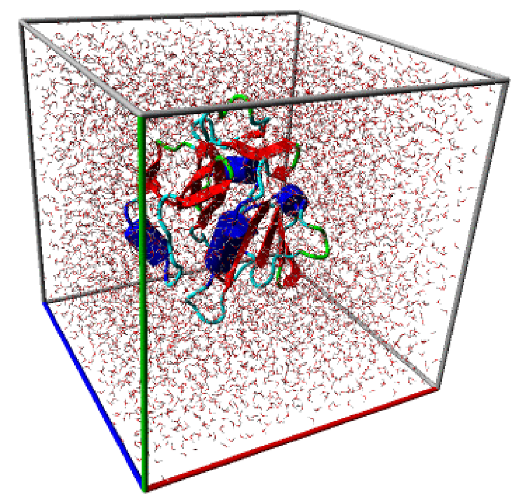

Fig. 2: A Molecular Dynamics simulation box containing the protein dihydrofolate reductase and water molecules as solvent (from [5]). The hundreds of millions of possible pair-wise interactions and resulting movements in such complex systems can only be approximated.

3. Aim

A number of chemical shift predictors exist which use a variety of techniques from artificial neural networks to protein homology [8-11]. A recent study using Random Forests non-linear regression suggest that there is room for improvement in all cases [12].

Our aim is twofold:

- Achieve better performance than current approaches

- Develop a fast-to-compute, easily differentiable function that, unlike existing predictors, can easily be implemented into MD simulations

4. Project Status

Based on data fitting between chemical shift data [13, 14] and protein structure [15] a preliminary version of the prediction algorithm has been developed. The main 'CamShift' equation consists of a linear combination of terms:

$$\delta = \delta_c + \delta_{dbb} + \delta_{dsc} + \delta_{dcb} + \delta_{hb} + \delta_{ar},$$

where the scalar δ on the left hand side is the chemical shift value for a given atom and the terms on the right depend on inter-atomic distances (δ_{dbb}, δ_{dsc}, δ_{dcb}), as well as capturing more complex hydrogen bonding and aromatic ring effects (δ_{hb}, δ_{ar}). δ_c is a constant.

At present, this equation was trained with data for two different atom types in a protein: Hα and Cα.

5. Preliminary Results

This first version of the predictor achieves a root mean square deviation (rmsd) between predicted and experimental chemical shifts of **0.28 ppm**[1] for Hα and **1.52 ppm** for Cα, which is a significant improvement to the results for an uninformed best guess (based on always predicting the mean of the sample) resulting in rmsds of **0.57 ppm** for Hα and **4.9 ppm** for Cα.

However, this is not yet as good as other predictors that report accuracies of up to **0.23 ppm** (Hα) and **0.98 ppm** (Cα) in some cases [11].

[1] ppm = parts per million, the unit of chemical shifts

6. Future Work

In short, future steps will be:

- Optimization and extension of terms for c.s. contributions
- Extension of the algorithm to work on other atom types
- Detailed comparison of predictive performance with that of existing predictors
- Implementation of CamShift into Molecular Dynamics package
- Applying new implementation to protein folding simulations

Acknowledgements

I would like to thank my supervisor, Dr. Michele Vendruscolo for his continuous support, and various members of the Dobson and Vendruscolo groups as well as Prof. Martin Zacharias from International University Bremen, for helpful discussion and suggestions. I am grateful for funding from Microsoft Research Cambridge.

References

[1] Fersht AR (1998) ***Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding***, *W. H. Freeman*, 3rd Rev. Ed., ISBN 0716732688

[2] Dobson CM (2003) ***Protein folding and misfolding***, *Nature* 426: 884-890

[3] Selkoe DJ (2003) ***Folding proteins in fatal ways***, *Nature* 426: 900-904

[4] Vendruscolo M, Dobson CM (2005) ***Towards complete descriptions of the free-energy landscapes of proteins***, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 363: 433-452

[5] *http://www.yasara.org/benchmarks.htm*

[6] Güntert P (1998) **Structure calculation of biological macromolecules from NMR data**, *Quarterly Reviews of Biophysics* 31(2): 145-237

[7] Levit MH (2001) Chapter 7.7, In: ***Spin Dynamics – Basics of Nuclear Magnetic Resonance***, *John Wiley & Sons Ltd*, ISBN 0471489220

[8] Wishart DS, Watson MS, Boyko RF, Sykes BD (1997) ***Automated ^{1}H and ^{13}C Chemical Shift Prediction Using the BioMagResBank,*** *Journal of Biomolecular NMR* 10, 329-336

[9] Xu XP, Case DA (2001) ***Automated prediction of ^{15}N, ^{13}Ca, ^{13}Cb and ^{13}C' chemical shifts in proteins using a density functional database***, *Journal of Biomolecular NMR* 21: 321-333

[10] Meiler J (2003) ***PROSHIFT: protein chemical shift prediction using artificial neural networks,*** *Journal of Biomolecular NMR*, 26(1):25-37

[11] Neal S, Nip AM, Zhang H, Wishart DS (2003) ***Rapid and accurate calculation of protein 1H, 13C and 15N chemical shift,*** *Journal of Biomolecular NMR*, 26:215-240.

[12] Arun K, Langmead CJ (2005) ***Structure based chemical shift prediction using Random Forests non-linear regression,*** *Carnegie Mellon University School of Computer Science Technical Report CMU-CS-05-163*

[13] Cornilescu G, Delaglio F, Bax A (1999) ***Protein backbone angle restraints from searching a database for chemical shift and sequence homology***, *Journal of Biomolecular NMR* 13: 289-302

[14] Zhang H, Neal S, Wishart D (2003) ***RefDB: A database of uniformly referenced protein chemical shifts***, *Journal of Biomolecular NMR* 25: 173-195

[15] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) ***The Protein Data Bank***, *Nucleic Acids Research* 28: 235-242

The foreground graphic shows the side-chain of the natural amino acid tyrosine