Solubility as an evolutionary constraint

Hugh P. Shanahan* & J.M. Thornton

European Bioinformatics Institute, Cambridge

1 Introduction

It has become increasingly apparent that the solubility of a protein plays a crucial rôle in its stability and function (Dobson, 2002). Proteins must, and do, exhibit a huge range of solubilities in water. Structural proteins need to be insoluble while active proteins, such as enzymes, which may occur at high concentrations *in vivo*, must be highly soluble. This suggests that although many residues on the surface are not closely packed or critical for function, they are nevertheless constrained during evolution for reasons of solubility. A theoretical analysis of protein surface composition could give us some insight into this behaviour. Unfortunately, little is understood, from a theoretical standpoint, of the relationship between the hydrophobic content, structure, and solubility of the protein. Predicting the solubility of a protein from its structure (let alone the original sequence) lies beyond the limits of present physical/chemical techniques.

There are two clear experimental observations. In the first instance, as noted by Schein (2001) and Sim and Sim (1999), the solubility of a protein is predicated upon an excess of polar atoms on the surface of the protein in question (namely, an absence of large hydrophobic patches). Furthermore, the solvation of the surface of a protein is a collective phenomenon. This is borne out from high resolution X-ray crystallography (Nakasako 2001) where the position of surface waters can be placed reasonably accurately.

In the light of the above observation, we hypothesize that residues on the surface are not individually constrained by solubility requirements, as opposed to the core, where a single mutation can affect packing or a key folding event. Instead, we suggest that solubility is a cooperative phenomenon best understood by considering local groups of residues or surface patches. With this in mind, we present a study of the hydrophobic character of the surfaces of a set of monomers and ask if we observe any constraints on them. We attempt to characterize the nature of patches on these proteins amongst homologous proteins. In particular we ask, in the first instance, if individual mutations occur independently of neighbouring residues. We do so by comparing the distribution of resulting average hydrophobicities for each patch against a distribution of average hydrophobicities generated from randomized sequences which satisfy the variation of residues at each sequence site. Furthermore, we examine if in suppressing large hydrophobic patches on solvent accessible surfaces of proteins the relative placement of individual atoms is also constrained. We do this by comparing the average hydrophobicities of observed patches with the averaged hydrophobicities of patches from randomised surfaces, where in this case the atom types have been swapped around.

2 Methods and Results

2.1 Hydrophilic constraint in the evolution of patches

To characterize the local environment on the surface of a protein, we define the Residue Hydrophobic Density (RHD) for a residue a as

$$RHD_a = \frac{1}{N_r} \sum_b H_b, b \in \mathcal{N}_a^R \ . \tag{1}$$

The residue hydrophobicity scale H_b is similar to that defined by Fauchere and Pliska (1983), which was determined using octanol/water distribution coefficients described by Yunger and Cramer (1981) and Cornette *et al.* (1987). It should be noted that we multiplied this scale by -1, in order to maintain the convention that increasing the horizontal scale increases the hydrophobicity, and multiplied it by 100 so that we could use integer based arithmetic. In order to define the neighbourhood $\mathcal{N}_a^R$, we defined the centroid of the accessible atoms of a particular residue as the average of their displacements and then using the resulting centroids and a cut-off of 12 Å.

This parameter is computed for all possible patches on the surfaces of 28 monomeric soluble protein collated by Ponstingl *et al.* (2000) (a larger set of monomeric proteins is available, but this subset gives us a sufficient number of homologues). Close homologues are determined for each protein using PSI-Blast. We assume that the residues in the homologues have the a similar spatial relationship with each other and compute this parameter for equivalent patches on each of the homologues. Hence we can derive a distribution of the RHS for each patch over each set of homologues.

In order to create a control, for each protein and their homologues, we generate randomised sequences which, at each site, have the same possible set of residues as the observed sequences.

In figure 1 we plot a distribution of the means of these observed distributions and the equivalent randomised ones. We note that the distribution of the observed means is noticeably more hydrophilic than the distribution of the randomised means (although the hydrophobic tails are very similar).

2.2 Constraints on atomic placement

To characterize the local environment on the surface of a protein, we define the Atomic Hydrophobic Density (AHD) in the similar fashion as equation 1, namely

$$AHD_i = \frac{1}{N_i} \sum_j h_j, j \in \mathcal{N}_i \ . \tag{2}$$

While it is possible to construct a more complicated neighbourhood, we employ a simple definition of the neighbourhood where the neighbouring atom j must lie in a sphere (of radius 6 Å) around atom i, that is

$$j \in \mathcal{N}_i \Leftrightarrow |\vec{r_j} - \vec{r_i}| < R_0 \tag{3}$$

where $\vec{r_j}$ is the displacement of atom j and we take $R_0 = 6$ Å. In order to avoid using a scoring system based directly on its environment we define h_j as the hydrophobicity of atom j, using a scale defined by Eisenberg and McLachlan (1985), where the heavy atoms are divided into 5 classes C, N/O, O^-, N^+ and S and takes the values, respectively 16, -6, -24, -50 and 21. This is based on an estimate of the change of the free energy of these atom types when solvated.

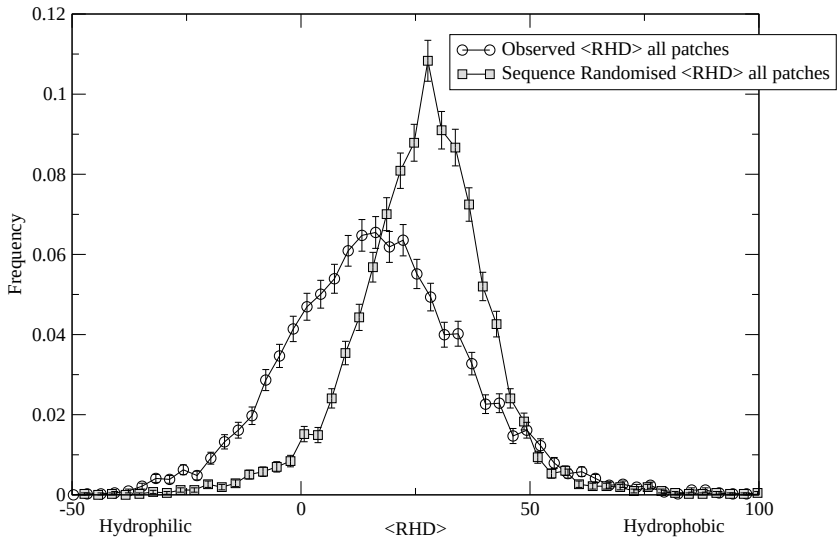

Figure 1: Histograms of the means of the RHD for all patches of all proteins, determined from observed and randomized protein sequences.

This parameter can be computed for all possible patches of 98 monomeric proteins from Ponstingl *et al.* (2000).

In measuring the distribution of the AHD outlined above, we must generate a set of randomised surfaces to provide a set of measurements to compare with. The resulting surface of atoms will not necessarily satisfy any stereochemistry. However, we can construct a randomisation algorithm that ensures that no two polar atoms are in contact (less than 2 Å separation) with each other (equivalent to being covalently bonded). We employ the following algorithm. For each atom position, we determine its possible neighbours that lie within 2 Å of it. We randomly pick an unlabelled atom position and label it as a polar atom and label all its neighbours as hydrophobic. This is repeated until all the atom positions are labelled. The individual atom types on the Eisenberg-Mclachlan scale are allocated randomly amongst the labelled atom types so as to reflect their individual total frequencies for that protein.

In figure 2 we plot the AHD of all possible observed and randomised patches. We see no significant difference between the distributions.

3 Conclusions

We examined how surface patches, defined at a residue-based level have evolved and observed that such proteins evolve to maintain their average hydrophilic content. In general, the hydrophilicity of the patch is more important than the hydrophilicity of an individual surface residue (and indeed than that of pairs of residues). On the other hand, we see no evidence at an atomic level that the position of the atoms are constrained, apart from the very simple requirement that polar atoms are greater than 2 Åfrom each other. This is possible because of the clever design of amino acids, where even the most hydrophobic amino acid will still have a polar atom (at least in its backbone) to interrupt a hydrophobic patch.

This indicates a much softer mode of evolution to the core, where individual residues are highly conserved. This plasticity would allow the evolution of, for example, enzymatic function or protein-protein interfaces. Future studies of the protein surfaces should then focus on such patches

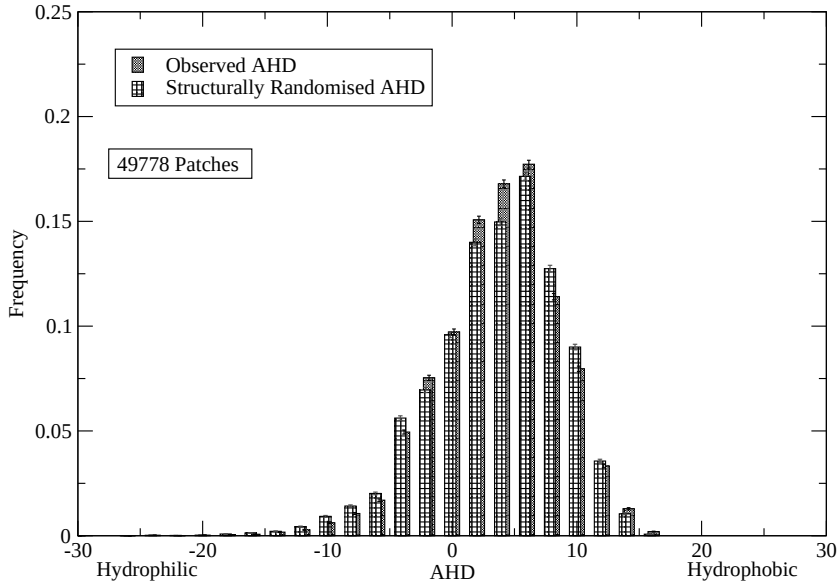

Figure 2: A histogram of the atomic hydrophobic density for all patches on all proteins from the observed and randomized distributions.

References

Dobson, C. (2002). Protein-misfolding diseases: Getting out of shape. *Nature*, 418:729–730.

Schein, C. H. (2001). Solubility as a function of protein structure and solvent components. Edited by Hardin, C. *et al.*. *Cloning Gene Expression and Protein Purification*. Oxford University Press.

Sim, J. and Sim, T.-S. (1999). Amino acid substitutions affecting protein solubility: high level expression of *streptomyces clavuligerus* isopenicillin n synthase in *escherichia coli*. *J. Mol. Cat B: Enzym.*, 6:133–143 and references therein.

Fauchere, J. L. and Pliska, V. (1983). Hydrophobic parameters-pi of amino-acid side-chains from the partitioning of n-acetyl-amino-acid amides. *Euro. J. Med. Chem.*, 18:369–375.

Yunger, L. M. and Cramer, R. D. (1981). Measurement and correlation of partition coefficients of polar amino acids. *Mol. Pharmacol.*, 20:602–608.

Cornette, J. L., Cease, K., Margalit, H., Spouge, J. L., Berzofsky, J., and Delisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.*, 195:659–685.

Ponstingl, H., Henrick, K., and Thornton, J. M. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, 41:47–57.

Eisenberg, D. and McLachlan, A.D. (1986), Solvation energy in protein folding and binding. *Nature*, **319**, 199-203.