Supplementary Information for: "Thermodynamics of stacking interactions in proteins"

Simone Marsili, Riccardo Chelli, Vincenzo Schettino, and Piero Procacci

Dipartimento di Chimica, Università di Firenze, Via della Lastruccia 3, I-50019 Sesto Fiorentino, Italy;
European Laboratory for Nonlinear Spectroscopy (LENS), Via Nello Carrara 1, I-50019 Sesto Fiorentino, Italy

I. EXCESS FREE ENERGIES FOR RESIDUE PAIR ARRANGEMENTS FROM DIFFERENT DATABASES OF PROTEIN STRUCTURES

Many studies similar to ours, i.e. aimed at quantitatively determining free energy differences, are usually carried out using public domain databases of non homologous proteins, while our database contains proteins with homology as large as 95%. However, provided that the database is sufficiently large, the results obtained by including or not including homologous proteins are expected to be quite similar as far as the free energy differences of residue pair arrangements or in general the contact potentials of residues are concerned. In some sense the "ideal" protein database is expected to behave as if the thermodynamic limit has been reached, which simply means in the limit of a large number of proteins.

Hence, the basic question could be: "is the current available protein database (the well known Protein Data Bank, PDB) representative of such a thermodynamic limit?" Of course, a definitive answer to this question is difficult to be given. Less ambitiously, one could get insightful information on this fundamental issue by comparing the behavior of protein databases which, in turn, include or not include homologous proteins. This is the issue that we will address in the present short report.

To this aim, we have performed the calculation of the potential of mean force (PMF) $w'_{RR'}(r,\theta)$ (see Eq. 2.8 of the manuscript) for all possible residue pairs using two of the most popular "non redundant" databases: the FSSP database[1] and the SCOP database[2]. As one can see from a selection of PMFs (Figure 1), the differences between PDB (our), FSSP, and SCOP databases are not significant. The three databases agree not only in the most relevant features, but also in the finer details. It should be also noted that the differences between FSSP and SCOP are of the same order of the differences between SCOP and PDB, or FSSP and PDB.

In order to evaluate the overall differences among the databases with respect to the stacked excess free energies, in Figure 2 we show the PDB-FSSP, PDB-SCOP, and FSSP-SCOP correlation diagrams of the stacked excess free energy of all residue pairs (calculated according to Eq. 2.7 of the manuscript). The regression and correlation coefficients of the diagrams are reported in the figure. The correlation of PDB with both SCOP and FSSP is very satisfactory. In general, the correlation between the PDB database and the SCOP and FSSP databases gets worse for high stacking excess free energies, i.e. for less populated stacked structures (see Figure 2A). The same holds true for the correlation between FSSP and SCOP (see Figure 2B). Given the excellent agreement among the databases for low excess free energy, it can be therefore safely stated that, for high excess free energies, the most reliable results are those provided by the largest database, i.e. the PDB one. This last statement is supported also by the fact that the FSSP-SCOP correlation is worse than the PDB-SCOP and PDB-FSSP ones (see regression and correlation coefficients in Figure 2).

References and notes

1 L. Holm, C. Ouzounis, C. Sander, G. Tuparev, G. Vriend, *Protein Sci.*, 1992, **1**, 1691-1698.
The current release of the FSSP protein database (25 November 2004; Web site: www.ebi.ac.uk/dali/fssp) has 2860 entries. From this set of non-homologous proteins, we have eliminated the DNA-protein complexes finally obtaining a database with 2434 proteins.

2 A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J. Mol. Biol.*, 1995, **247**, 536-540.
The current release of the SCOP protein database (1.73 release, November 2007; Web site: scop.mrc-lmb.cam.ac.uk/scop) has 3751 entries. From this set of non-homologous proteins, we have eliminated the DNA-protein complexes finally obtaining a database with 3043 proteins.

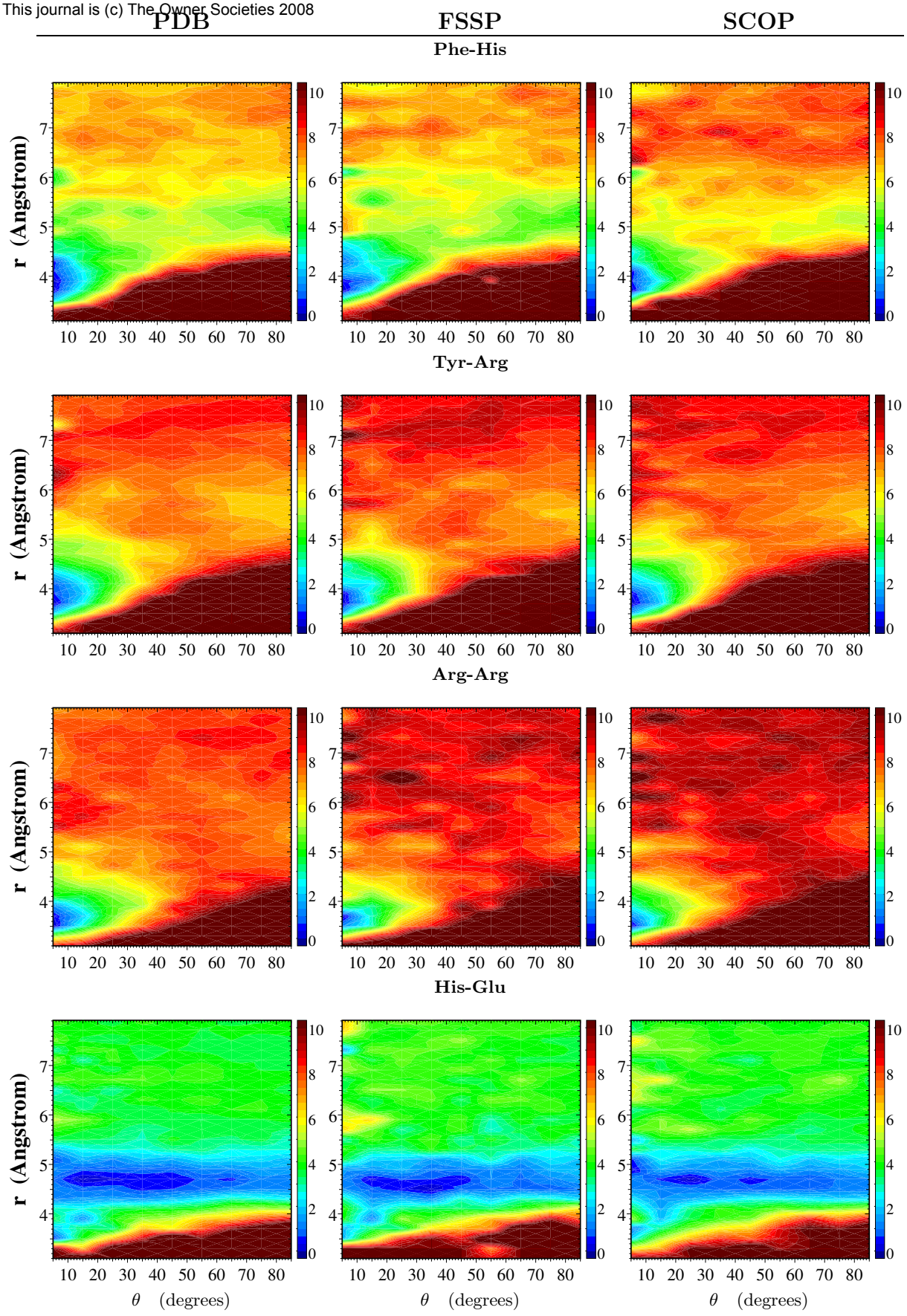

FIG. 1: Selected $w'_{RR'}(r,\theta)$ PMFs using the PDB, FSSP and SCOP protein databases.

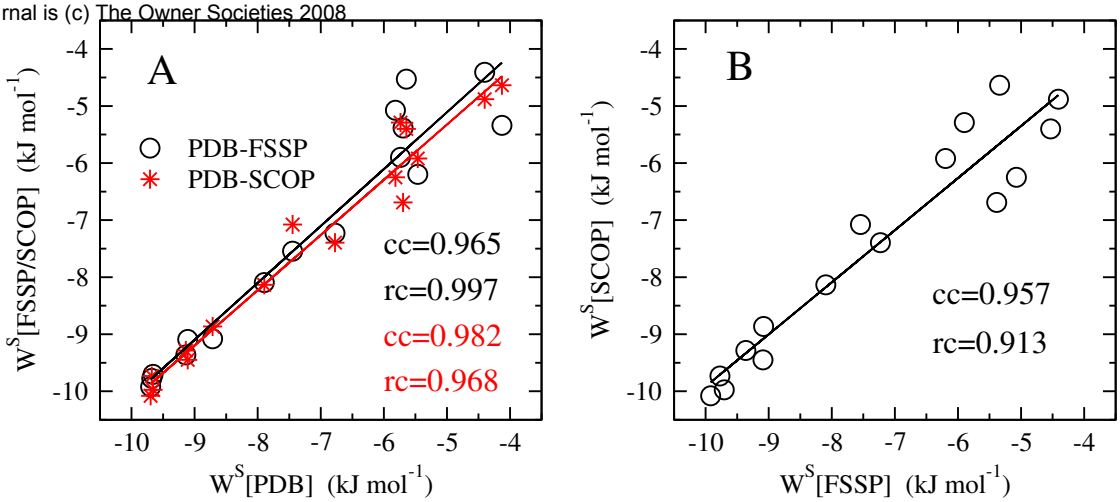

FIG. 2: PDB-FSSP, PDB-SCOP (panel **A**) and FSSP-SCOP (panel **B**) correlation diagrams of the stacked excess free energies for all residue pairs. The lines represent the regression fits for the diagrams whose correlation and regression coefficients (*cc* and *rc*, respectively) are also reported in the panels.