Journal of Computer-Aided Molecular Design, **15:** 721–740, 2001.
KLUWER/ESCOM

Implicit solvation in the self-consistent mean field theory method: sidechain modelling and prediction of folding free energies of protein mutants

Joaquim Mendes†, António M. Baptista, Maria Arménia Carrondo & Cláudio M. Soares*
Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Av. da República, Apartado 127, 2781-901 Oeiras, Portugal; †Present address: European Molecular Biology Laboratories (EMBL), Meyerhofstrasse 1, D-69117 Heidelberg, Germany

Received 27 December 2000; accepted 29 May 2001

Key words: atomic solvation parameter, flexible rotamer model, folding free energy, protein design, sidechain prediction, solvent accessible surface area

Summary

The Atomic Solvation Parameter (ASP) model is one of the simplest models of solvation, in which the solvation free energy of a molecule is proportional to the solvent accessible surface area (SAS) of its atoms. However, until now this model had not been incorporated into the Self-Consistent Mean Field Theory (SCMFT) method for modelling sidechain conformations in proteins. The reason for this is that SAS is a many-body quantity and, thus, it is not obvious how to define it within the Mean Field (MF) framework, where multiple copies of each sidechain exist simultaneously. Here, we present a method for incorporating an SAS-based potential, such as the ASP model, into SCMFT. The theory on which the method is based is exact within the MF framework, that is, it does not depend on a pairwise or any other approximation of SAS. Therefore, SAS can be calculated to arbitrary accuracy. The method is computationally very efficient: only 7.6% slower on average than the method without solvation. We applied the method to the prediction of sidechain conformation, using as a test set high-quality solution structures of 11 proteins. Solvation was found to substantially improve the prediction accuracy of well-defined surface sidechains. We also investigated whether the methodology can be applied to prediction of folding free energies of protein mutants, using a set of barnase mutants. For apolar mutants, the modest correlation observed between calculated and observed folding free energies without solvation improved substantially when solvation was included, allowing the prediction of trends in the folding free energies of this type of mutants. For polar mutants, correlation was not significant even with solvation. Several other factors also responsible for the correlation were identified and analysed. From this analysis, future directions for applying and improving the present methodology are discussed.

Introduction

The solvent environment associated with biological molecules has profound influences on the structure, dynamics and thermodynamics of such systems in their active, or native, state [1]. It is also thought to play a dominant role in the folding process whereby biological macromolecules acquire their native structure [2].

The effect of the solvent environment on the properties of a molecule can be quantified by the solvation free energy, ΔG^{sol}, which corresponds to the free energy change associated with the transfer of the molecule from vacuum to the solvent [3]. This free energy is a complex quantity that represents not only the interactions between the solute and the solvent molecules, but also the changes that occur in the interactions of the solvent molecules among themselves resulting from introducing the solute into the solvent, such changes being not only enthalpic but also entropic

*To whom correspondence should be addressed. E-mail: claudio@itqb.unl.pt

[3,4]. The solvation free energy may be considered to comprise three components:

$$\Delta G^{\text{sol}} = \Delta G^{\text{elec}} + \Delta G^{\text{vdW}} + \Delta G^{\text{cav}}, \quad (1)$$

where ΔG^{elec} is an electrostatic component that quantifies the electrostatic interaction of the solute molecule with the solvent molecules and the changes in the electrostatic interactions of the solvent molecules among themselves, ΔG^{vdW} is a van der Waals component that quantifies the van der Waals interactions of the solute molecule with the solvent molecules, and ΔG^{cav} is a component that represents the free energy required to form the solute cavity within the solvent. This last term is associated only with the solvent and results from the reorganisation of the solvent molecules around the solute cavity and the work done against the solvent pressure to create the solute cavity [3]. When the solvent is water, as is usually the case for biomolecules, it is predominantly entropic at room temperature [2].

A substantial amount of theoretical effort has been directed toward providing models and theories capable of describing the solvation effects observed experimentally. The most general means of modelling solvent is to include solvent atoms explicitly [1]. This approach models the three components of solvation simultaneously, and when incorporated in molecular dynamics, includes both the equilibrium and dynamic effects of solvation. It also models microscopic structural details of non-bulk water in the vicinity of the solute, such as water bridging. The major disadvantage of explicit solvent simulations is the large computational expense involved.

For systems at equilibrium, solvation can be modelled by an alternative approach employing macroscopic or semi-macroscopic methods, which are orders of magnitude faster than explicit solvent methods [1]. In this approach, the electrostatic component of solvation is modelled independently of the sum of the van der Waals and cavity components. The most commonly used methods for modelling the electrostatic component are the Poisson–Boltzmann methods [5] and the Langevin dipoles method [6]. Although these methods do not model the microscopic structural details of non-bulk water nor, in general, the dynamic effects of solvation, they do model the dielectric screening between charged groups of the solute that results from their interaction with the solvent. The sum of the van der Waals and cavity components is usually modelled as a linear function of the total solvent accessible surface area (SAS) of the solute [7], which is based on the linear dependence of solvation free energy on SAS observed for hydrocarbons [8–10].

Solvation can also be modelled approximately using the Atomic Solvation Parameters (ASP) approach [11], which is considerably faster than the methods described above. This approach assumes that the solvation free energy of the solute can be decomposed into a sum of atomic free energy components that are proportional to the solvent exposure of the atom:

$$\Delta G^{\text{sol}} = \sum_a \sigma_a \, \text{SAS}_a, \quad (2)$$

where SAS_a is the SAS of atom a, the proportionality constant σ_a is the atomic solvation parameter that depends on the chemical nature of this atom, and where the summation runs over all atoms of the solute. Although this method effectively models the relative affinity of different chemical groups for solvent, it does not model the dielectric screening between charged groups of the solute. Therefore, a screening function, such as a distance-dependent dielectric constant, should also be included to more adequately model solvation [1].

The ASP model has been applied to a vast range of problems including molecular dynamics simulations [12], homology modelling [13], modelling of sidechain conformation [14–17], refinement of experimental protein structure models [18] and prediction of free energies of mutation [19]. Of particular interest here is its application to the modelling of sidechain conformation. In spite of it having been shown that inclusion of an ASP model for solvation does improve sidechain prediction in homology modelling [14–16], this application has been limited to only a few of the available sidechain modelling methods. These methods are all rotamer-based and perform a combinatorial rotamer search [14, 15] or a Monte Carlo optimisation [16]. In either case, a particular rotamer combination is generated at each step of the algorithm and the solvation free energy of the *single* protein conformation corresponding to this rotamer combination can be calculated by directly applying expression (2). Application of the ASP model to other commonly used sidechain modelling methods, such as the Dead-End Elimination method [20] and the Self-Consistent Mean Field Theory method (SCMFT) [21], has been hampered by the fact that such a direct application of expression (2) is not possible. The reason for this in the Dead-End Elimination method, on one hand, is that this method requires that all energy components be decomposable into pairwise terms, that is, terms

that depend only on a pair of atoms. However, this is not possible for the solvation free energy function described by the ASP model, because the SAS of an atom is a many-body quantity that cannot be decomposed into pairwise terms. Efforts have recently been made to develop an approximate pairwise decomposition of SAS in an attempt to overcome this problem [22]. In the SCMFT method, on the other hand, the difficulty in applying the ASP model resides in the fact that all possible sidechain rotamers have to be considered simultaneously at a given residue position and, therefore, it is not obvious how to define the SAS of a particular atom. Until now, no implementation of the ASP model in SCMFT has been described [23].

In this work we develop a theoretical method for implementing the ASP model into the SCMFT method [21]. The proposed method does not depend on a pairwise decomposition of SAS and, therefore, the accuracy of the calculated SAS is dictated only by the accuracy of the method used to calculate surface area. We show that, within the MF framework, the usual SAS of an atom for a single protein conformation is substituted by an average SAS over all possible rotamer combinations. In each cycle of the SCMFT algorithm, this average SAS correctly takes into consideration the backbone and all rotamers at all residues weighted by their respective probabilities. Moreover, contrary to the combinatorial search and Monte Carlo methods described above, where calculation of SAS in each step of the algorithm can become computationally very expensive, the calculation of SAS in each cycle of SCMFT is computationally cheap because extensive, but rigorous, simplification of the expressions involved is possible, and because SCMFT converges within very few cycles. Thus, for the set of 11 proteins on which sidechain prediction was assessed (see below), the method with solvation was only 7.6% slower on average than the method without solvation.

We applied the methodology developed in this work to two distinct problems. The first was a straightforward application of the SCMFT method to modelling of sidechain conformation, where we analysed the effect of solvation on the accuracy with which the conformation of well-defined residues in the solution structures of 11 proteins was predicted. The second was an extension of the SCMFT method to prediction of folding free energies of protein mutants. The basis of this application was that MF Theory provides a simplified free energy function that includes both the energetic and entropic consequences of sidechain conformational multiplicity, allowing the SCMFT method to be extended directly to the prediction of folding free energies. This is a particular advantage of the method over other sidechain modelling methods frequently applied in this area, which assume that the free energy can be approximated by the potential energy of a single protein conformation. The aim of the present study was to investigate if the MF free energy function is sufficiently accurate to allow prediction of folding free energies when a rotamer description of sidechain conformational space is used. To this end, the methodology was applied to all apolar and polar mutants of barnase published in an extensive study performed in the Fersht group [24–30]. We compared the performance of the method with and without solvation.

Theory

The system

The system of interest to a sidechain modelling method including solvation is a protein with a rigid backbone and N flexible sidechains, plus the surrounding solvent. In rotamer-based methods, a sidechain i may exist only in a discrete set of conformational states r_i called rotamers. The global conformational state of the protein can, therefore, be represented as $\mathbf{r} = (r_1, r_2, \ldots, r_N)$. Since solvation is treated implicitly in what follows, the configurational states of the solvent do not appear explicitly and, therefore, need not be defined here.

All thermodynamic properties of the protein-solvent system can be derived from its free energy[1]:

$$G = \left\langle E \right\rangle + \left\langle \Delta G^{\mathrm{sol}} \right\rangle - TS, \tag{3}$$

where $\langle E \rangle$, $\left\langle \Delta G^{\mathrm{sol}} \right\rangle$ and S are, respectively, the average intramolecular potential energy of the protein, its average solvation free energy and its conformational entropy, and T is the absolute temperature. The average of any microscopic physical quantity X of the protein is given by [31]:

$$\langle X \rangle = \sum_{\mathbf{r}} p(\mathbf{r}) X(\mathbf{r}), \tag{4}$$

where $p(\mathbf{r})$ is the probability of the global conformational state **r,** $X(\mathbf{r})$ is the value assumed by X in this

[1] G corresponds more precisely to the difference between the free energy of the protein-solvent system and the free energy of the pure solvent system.

conformational state, and where the summation in $\mathbf{r}$ runs over all possible rotamer combinations of the N amino acid sidechains of the protein. The entropy is defined by [31]:

$$S = -R\sum_{\mathbf{r}} p(\mathbf{r}) \ \ln \ p(\mathbf{r}), \tag{5}$$

where R is the ideal gas constant.

The mean field approximation

At equilibrium, the probability of a particular global conformational state $\mathbf{r}$ is given by the Boltzmann relation. Therefore, the free energy of the protein system at equilibrium can, in principle, be calculated by direct application of expressions (4) and (5). However, this direct approach is rarely computationally tractable since, even for a protein with a relatively small number of residues, the number of possible rotamer combinations over which the summations in these expressions are carried out is very large. One means of overcoming this combinatorial problem is to assume an approximate functional form for the global probability $p(\mathbf{r})$ described by the Mean Field relation:

$$p(\mathbf{r}) = \prod_{i}^{N} p_i(r_i), \tag{6}$$

where $p_i(r_i)$ is the probability of sidechain i existing in the rotamer state r_i. Under this approximation, the summations in expressions (4) and (5) reduce to summations with a much smaller number of terms, corresponding to the backbone and individual sidechains of the protein.

The free energy function

The functional forms assumed by $\langle E \rangle$ and S under the MF approximation have previously been derived [31, 32] and are given by:

$$\begin{aligned} \langle E \rangle = & \sum_{i}^{N}\sum_{r_i} p_i(r_i)\Big[U_i(r_i) \\ & + \tfrac{1}{2}\sum_{j\neq i}^{N}\sum_{r_j} p_j(r_j)U_{ij}(r_i, r_j)\Big] \\ = & \sum_{i}^{N}\sum_{r_i} p_i(r_i)\,\langle E_i\rangle_{r_i}\,, \end{aligned} \tag{7}$$

$$S = -R\sum_{i}^{N}\sum_{r_i} p_i(r_i)\ln p_i(r_i), \tag{8}$$

where $U_i(r_i)$ is the sum of the intrinsic potential energy of sidechain i when in rotamer state r_i and its potential energy of interaction with the backbone, and $U_{ij}(r_i,r_j)$ is the potential energy of interaction between sidechains i and j when in rotamer states r_i and r_j respectively. E_i can be interpreted as the contribution of sidechain i to the total energy of the system and $\langle E_i\rangle_{r_i}$ is its conditional average when the sidechain is fixed in rotamer state r_i, defined by an obvious extension of expression (4).

From definition (4) and the ASP model for the solvation free energy of a single global conformation $\mathbf{r}$ (expression (2)), it is always possible to decompose $\langle \Delta G^{\mathrm{sol}} \rangle$ into sidechain and backbone terms:

$$\begin{aligned} \langle \Delta G^{\mathrm{sol}} \rangle \ &= \ \sum_{i}^{N}\sum_{r_i} p_i(r_i)\langle \Delta G_i^{\mathrm{sol}}\rangle_{r_i} + \langle \Delta G_{bb}^{\mathrm{sol}}\rangle \\ &= \ \sum_{i}^{N}\sum_{r_i} p_i(r_i)\sum_{a\in i}\sigma_a\,\langle \mathrm{SAS}_a\rangle_{r_i} \\ &\quad + \sum_{a\in bb}\sigma_a\,\langle \mathrm{SAS}_a\rangle\,, \end{aligned} \tag{9}$$

where $\langle \Delta G_i^{\mathrm{sol}}\rangle_{r_i}$ is the average solvation free energy of sidechain i when fixed in rotamer state r_i, $\langle \Delta G_{bb}^{\mathrm{sol}}\rangle$ is the average solvation free energy of the backbone, $\langle \mathrm{SAS}_a\rangle_{r_i}$ is the conditional average SAS of atom a when sidechain i is fixed in rotamer state r_i, and $\langle \mathrm{SAS}_a\rangle$ is the average SAS of atom a. Therefore, derivation of the functional form assumed by $\langle \Delta G^{\mathrm{sol}}\rangle$ under the MF approximation reduces to the derivation of that assumed by $\langle \mathrm{SAS}_a\rangle$ and $\langle \mathrm{SAS}_a\rangle_{r_i}$ under this approximation. The SAS of an atom a is defined as the area of the solvation sphere of that atom that is not buried by (the solvation sphere of) any other atom $a' \neq a$ [7], where the solvation sphere of an atom is a sphere centred at the atom and of radius equal to the sum of the van der Waals radii of the atom and a solvent molecule. For the particular problem at hand, the SAS of an atom a in the conformational state $\mathbf{r}$ of the protein may be calculated as:

$$\mathrm{SAS}_a(\mathbf{r}) = \begin{cases} \int_{S_a} \alpha^{bb}(\mathbf{s}_a)\prod_{i}^{N}\alpha^{i}(\mathbf{s}_a, r_i)\,\mathrm{d}s_a \text{ for} \\ \text{for } a \in bb \\ \int_{S_a} \alpha^{bb}[\mathbf{s}_a(r_i)]\prod_{j}^{N}\alpha^{j}\big[\mathbf{s}_a(r_i), r_j\big]\,\mathrm{d}s_a \text{ for} \\ \text{for } a \in i \end{cases} \tag{10}$$

where $\mathbf{s}_a$ is a point on the surface of the solvation sphere of atom a defining the position of the surface element $\mathrm{d}s_a$; for a backbone atom, $\mathbf{s}_a$ is independent of the global conformational state $\mathbf{r}$, and for a sidechain atom, it depends only on the rotamer state of the sidechain in the global conformational state $\mathbf{r}$.

$\alpha^{bb}(\mathbf{s}_a)$ is a function that equals zero if $\mathbf{s}_a$ is buried by at least one backbone atom $a' \neq a$ and one if it is not. Similarly, $\alpha^i(\mathbf{s}_a, r_i)$ is a function that equals zero if $\mathbf{s}_a$ is buried by at least one atom $a' \neq a$ of sidechain i in rotamer state r_i and one if it is not. The functional forms assumed by $\langle \mathrm{SAS}_a \rangle$ and $\langle \mathrm{SAS}_a \rangle_{r_i}$ under the MF approximation can finally be derived from expression (10), the MF relation (6) and the definitions of average (expression (4)) and conditional average quantities, which after some rearrangement of terms yields:

$$\begin{aligned}
&\langle \mathrm{SAS}_a \rangle = \int_{S_a} \left[\alpha^{bb}(\mathbf{s}_a) \prod_i^N \sum_{r_i} p_i(r_i) \alpha^i(\mathbf{s}_a, r_i) \right] \mathrm{d}s_a \\
&\text{for } a \in bb, \\
&\langle \mathrm{SAS}_a \rangle_{r_i} = \int_{S_a} \Big\{ \alpha^{bb}\left[\mathbf{s}_a(r_i)\right] \alpha^i\left[\mathbf{s}_a(r_i), r_i\right] \\
&\prod_{j \neq i}^N \sum_{r_j} p_j(r_j) \alpha^j\left[\mathbf{s}_a(r_i), r_j\right] \Big\} \mathrm{d}s_a \quad \text{for } a \in i.
\end{aligned} \tag{11}$$

It is easy to show that the integrands of these expressions represent the probability of $\mathbf{s}_a$ not being buried by any atom in the protein.

Equilibrium probabilities

At this point the free energy of the protein system under the MF approximation cannot yet be calculated from expressions (7), (8), (9) and (11) because the sidechain probabilities $p_i(r_i)$ that appear in these expressions are undetermined. However, these probabilities can be determined in the MF framework for the system at equilibrium by taking into account that, at equilibrium, the free energy of the system is at a minimum. The set of probabilities that minimise G must simultaneously satisfy the constraint conditions:

$$\sum_{r_i} p_i(r_i) = 1 \quad \forall i. \tag{12}$$

The extrema of a function subject to constraints can be determined with the method of Lagrange multipliers, which in the present case leads to a set of conditions [31, 32]:

$$\begin{aligned}
\frac{\partial G}{\partial p_i(r_i)} + \lambda_i &= \frac{\partial \langle E \rangle}{\partial p_i(r_i)} + \frac{\partial \langle \Delta G^{\mathrm{sol}} \rangle}{\partial p_i(r_i)} \\
&\quad - T \frac{\partial S}{\partial p_i(r_i)} + \lambda_i = 0 \quad \forall i, r_i,
\end{aligned} \tag{13}$$

where λ_i is the undetermined Lagrange multiplier of sidechain i. The partial derivatives of $\langle E \rangle$ and S under the MF approximation have previously been derived [31,32] and are given by:

$$\begin{aligned}
\frac{\partial \langle E \rangle}{\partial p_i(r_i)} &= U_i(r_i) + \sum_{j \neq i}^N \sum_{r_j} p_j(r_j) U_{ij}(r_i, r_j) \\
&= \langle E_i^* \rangle_{r_i},
\end{aligned} \tag{14}$$

$$\frac{\partial S}{\partial p_i(r_i)} = -R \ \ln \ p_i(r_i) - R. \tag{15}$$

The partial derivatives of $\langle \Delta G^{\mathrm{sol}} \rangle$ under this approximation can be obtained by direct differentiation of expressions (9) and (11), which yields:

$$\frac{\partial \langle \Delta G^{\mathrm{sol}} \rangle}{\partial p_i(r_i)} = \left\langle \Delta G_i^{\mathrm{sol}} \right\rangle_{r_i} + \gamma_i - \left\langle \Delta G_i^{\mathrm{des}} \right\rangle_{r_i} \tag{16}$$

where $\langle \Delta G_i^{\mathrm{sol}} \rangle_{r_i}$ was defined above, γ_i is a constant for sidechain i independent of its rotamer state, and $\langle \Delta G_i^{\mathrm{des}} \rangle_{r_i}$ is the average *de*solvation free energy of the remainder of the protein system (that is, excluding sidechain i) that results from burial of surface area exclusively by atoms belonging to sidechain i when this sidechain is fixed in the rotamer state r_i and that is given by:

$$\begin{aligned}
\left\langle \Delta G_i^{\mathrm{des}} \right\rangle_{r_i} &= \sum_{j \neq i}^N \sum_{r_j} p_j(r_j) \sum_{a \in j} \sigma_a \left\langle {}^*\mathrm{SIS}_a^i \right\rangle_{r_i, r_j} \\
&\quad + \sum_{a \in bb} \sigma_a \left\langle {}^*\mathrm{SIS}_a^i \right\rangle_{r_i}
\end{aligned} \tag{17}$$

where

$$\begin{aligned}
\langle {}^*SIS_a^i \rangle_{r_i} = &\int_{S_a} \Big\{ \left[1 - \alpha^i(\mathbf{s}_a, r_i)\right] \alpha^{bb}(\mathbf{s}_a) \\
&\prod_{j \neq i}^N \sum_{r_j} p_j(r_j) \alpha^j(\mathbf{s}_a, r_j) \Big\} \mathrm{d}s_a \\
&\text{for } a \in bb, \\
\langle {}^*SIS_a^i \rangle_{r_i, r_j} = &\int_{S_a} \Big\{ \left\{1 - \alpha^i\left[\mathbf{s}_a(r_j), r_i\right]\right\} \\
&\alpha^{bb}\left[\mathbf{s}_a(r_j)\right] \alpha^j\left[\mathbf{s}_a(r_j), r_j\right] \\
&\prod_{k \neq i,j}^N \sum_{r_k} p_k(r_k) \alpha^k\left[\mathbf{s}_a(r_j), r_k\right] \Big\} \mathrm{d}s_a \\
&\text{for } a \in j \neq i,
\end{aligned} \tag{18}$$

${}^*\mathrm{SIS}_a^i$ being the part of the solvent inaccessible surface area (SIS) of atom a that results from exclusive burial by atoms belonging to sidechain i. It is easy to show that the integrands of expressions (18) represent the conditional probability of $\mathbf{s}_a$ being buried exclusively by at least one atom of sidechain i when fixed in the rotamer state r_i.

The functional form of the sidechain probabilities for the system at equilibrium can finally be obtained from expressions (13), (14), (15) and (16), using the constraint conditions (12) to eliminate λ_i and γ_i:

$$p_i(r_i)=\frac{\exp\left\{-\left[\langle E_i^*\rangle_{r_i}+\left\langle\Delta G_i^{\mathrm{sol}}\right\rangle_{r_i}-\left\langle\Delta G_i^{\mathrm{des}}\right\rangle_{r_i}\right]\Big/RT\right\}}{\sum_{r_i'}\exp\left\{-\left[\langle E_i^*\rangle_{r_i'}+\left\langle\Delta G_i^{\mathrm{sol}}\right\rangle_{r_i'}-\left\langle\Delta G_i^{\mathrm{des}}\right\rangle_{r_i'}\right]\Big/RT\right\}} \quad \forall i, r_i. \tag{19}$$

The desolvation free energy term, $\left\langle\Delta G_i^{\mathrm{des}}\right\rangle_{r_i}$, has been overlooked in other methodologies for including solvation in the SCMFT method [33]. However, it arises naturally in the present methodology and is clearly necessary whatever the solvation model incorporated in SCMFT. The importance of this term in determining the probability of sidechain i existing in rotamer state r_i can be appreciated by considering the mutual burial of surface area of two atoms, one belonging the sidechain i and the other to any other part of the protein. If the atom of sidechain i is hydrophobic and the other atom hydrophilic, the mutual burial of surface area of the two atoms will lead to a favourable decrease in the average solvation free energy of sidechain i but to an unfavourable increase in the average solvation free energy of the remainder of the system. Clearly, both of these effects must be taken into account in the sidechain probability function. The favourable decrease in average solvation free energy of the sidechain in rotamer state r_i is taken into account in $\left\langle\Delta G_i^{\mathrm{sol}}\right\rangle_{r_i}$. The unfavourable increase in the average solvation free energy of the remainder of the system when the sidechain is in rotamer state r_i is accounted for in the $-\left\langle\Delta G_i^{\mathrm{des}}\right\rangle_{r_i}$ term. A similar reasoning applies to any other combination of hydrophobic and hydrophilic atoms, and to any other solvation model.

Materials and methods

SCMFT methodology

In all calculations of the present work, the extended amino214 rotamer library [32], based on the amino214 rotamer library of Tufféry and co-workers (Tufféry *et al.*, 1997), was used for non-Pro residue types, and the set of rotamers described in [34] was used for Pro. Calculations were performed with and without solvation. The calculations without solvation were described in detail in [32, 34]. The calculations with solvation were in all aspects identical except that the MF free energy function and the equilibrium probability functions of the non-solvated system (Expressions (3) and (6), respectively, in [32]) were substituted by the corresponding functions of the solvated system (Expressions (3) and (19), respectively, derived in this work). The details of these calculations will be described in what follows.

Convergence algorithm

Similarly to what was previously discussed for the MF sidechain probabilities of the non-solvated system at equilibrium [21,32], $\langle E_i^*\rangle_{r_i}$, $\left\langle\Delta G_i^{\mathrm{sol}}\right\rangle_{r_i}$ and $\left\langle\Delta G_i^{\mathrm{des}}\right\rangle_{r_i}$, that appear in expression (19) defining the sidechain probabilities of the solvated system at equilibrium, are functions of the probabilities of all sidechains of the protein in all possible rotamer states. Therefore, the complete set of sidechain probabilities (corresponding to all sidechains in the protein in all possible rotamer states) constitutes a system of simultaneous non-linear equations in which the sidechain probabilities are the unknowns. This system of equations can be solved by the same self-consistent iterative method that was previously described for the non-solvated system [21], which converges to a minimum in the free energy function G. Thus, in the present work, converged values of the equilibrium sidechain probabilities for both systems were obtained with the self-consistent iterative method developed by Koehl and Delarue [21], implemented as in [32, 34] and incorporating the multiple run random initialisation (MRRI) protocol described in [32]. Each iteration of this method consists of two steps. In the first step, the value of the average energy of each rotamer ($\langle E_i^*\rangle_{r_i}+\left\langle\Delta G_i^{\mathrm{sol}}\right\rangle_{r_i}-\left\langle\Delta G_i^{\mathrm{des}}\right\rangle_{r_i}$) is updated by inserting the rotamer probabilities calculated in the previous iteration into expressions (14), (9) and (17), as well as into the numerical approximations for $\langle \mathrm{SAS}_a\rangle_{r_i}$, $\left\langle {}^*\mathrm{SIS}_a^{\mathrm{i}}\right\rangle_{r_i}$ and $\left\langle {}^*\mathrm{SIS}_a^{\mathrm{i}}\right\rangle_{r_i,r_j}$ given below in expression (20). In the second step of the iteration, the rotamer probabilities are updated by inserting the average rotamer energies calculated in the first step into expression (19). The MRRI protocol consisted of 200 runs of the iterative procedure. For each run of this protocol, it is necessary to evaluate the total MF free energy of the system using expression (3). This is performed by inserting the converged energies obtained at the end of the run into expressions (7), (8) and (9), as well as into the numerical approximations for $\langle \mathrm{SAS}_a\rangle_{r_i}$ and $\langle \mathrm{SAS}_a\rangle$ given below in expression (20).

Calculation of intramolecular potential energies

The intramolecular potential energy terms $U_i(r_i)$ and $U_{ij}(r_i, r_j)$, required to calculate $\langle E_i \rangle_{r_i}$ that appears in the MF free energy function and $\langle E_i^* \rangle_{r_i}$ that appears in the equilibrium sidechain probability functions of both the solvated and non-solvated systems, were calculated with the flexible rotamer model (FRM) [34]. Subrotamers were generated as described in [34] with flexibility both in bond angles and torsion angles and using a rejection threshold of three standard deviations. Effective rotamer energies were calculated with a sample of 1000 subrotamer pairs. The molecular mechanics potential energy function used in these calculations was the standard potential energy function used in GROMOS [35] incorporating the distance-dependent dielectric constant of Solmajer and Mehler [36, 37], as described in [34]; the distance-dependent dielectric constant was used to model dielectric screening, which is absent in the ASP model, as discussed in the introduction. All parameters were from the GROMOS force field with explicit aromatic and polar hydrogen atoms [35, 38].

Calculation of average SASs and average SISs

The average SASs of expressions (11) and the average SISs of expressions (18), required to calculate $\langle \Delta G^{\mathrm{sol}} \rangle$, $\langle \Delta G_i^{\mathrm{sol}} \rangle_{r_i}$ and $\langle \Delta G_i^{\mathrm{des}} \rangle_{r_i}$ that appear in the MF free energy function and in the equilibrium sidechain probability functions of the solvated system, were calculated with the classic rigid rotamer model (RRM) [34] (corresponding to a single conformation for each rotamer) using the following numerical approximations:

$$\langle \mathrm{SAS}_a \rangle = \sum_{s_a} \left[\alpha^{bb}(\mathbf{s}_a) \prod_i \sum_{r_i} p_i(r_i) \alpha^i(\mathbf{s}_a, r_i) \right] \Delta s_a \quad \text{for } a \in bb,$$

$$\langle \mathrm{SAS}_a \rangle_{r_i} = \sum_{s_a} \left\{ \alpha^{bb}[\mathbf{s}_a(r_i)]\ \alpha^i[\mathbf{s}_a(r_i), r_i] \prod_{j \neq i}^{N} \sum_{r_j} p_j(r_j) \alpha^j[\mathbf{s}_a(r_i), r_j] \right\} \Delta s_a \quad \text{for } a \in i,$$

$$\langle {}^*\mathrm{SIS}_a^i \rangle_{r_i} = \sum_{s_a} \left[\left[1 - \alpha^i(\mathbf{s}_a, r_i) \right] \alpha^{bb}(\mathbf{s}_a) \prod_{j \neq i}^{N} \sum_{r_j} p_j(r_j) \alpha^j(\mathbf{s}_a, r_j) \right] \Delta s_a \quad \text{for } a \in bb,$$

$$\langle {}^*\mathrm{SIS}_a^i \rangle_{r_i, r_j} = \sum_{s_a} \left\{ \left\{ 1 - \alpha^i[\mathbf{s}_a(r_j), r_i] \right\} \alpha^{bb}[\mathbf{s}_a(r_j)]\, \alpha^j[\mathbf{s}_a(r_j), r_j] \prod_{k \neq i,j} \sum_{r_k} p_k(r_k) \alpha^k[\mathbf{s}_a(r_j), r_k] \right\} \Delta s_a \quad \text{for } a \in j \neq i, \tag{20}$$

where the summations are over m finite surface elements s_a, each centred at $\mathbf{s}_a$ and of area Δs_a. An adaptation of the method of Shrake and Rupley [39] was used to compute these expressions. In this method, m dots are distributed uniformly on the surface of the solvated sphere of the atom, each dot representing the centre $\mathbf{s}_a$ of a surface element and all surface elements being assumed to have the same area $\Delta s_a = \Delta s = S_a / m$. In the adaptation of the method used here, m dots were distributed uniformly on each backbone atom and on each sidechain atom in each of the rotamer states of the sidechain. The summations in expressions (20) are carried out over a particular subset of these dots: the summations for $\langle \mathrm{SAS}_a \rangle$ and $\langle {}^*\mathrm{SIS}_a^i \rangle_{r_i}$ are carried out over the dots on backbone atom a, that for $\langle \mathrm{SAS}_a \rangle_{r_i}$ over the dots on atom a of sidechain i in rotamer state r_i, and that for $\langle {}^*\mathrm{SIS}_a^i \rangle_{r_i, r_j}$ over the dots on atom a of sidechain j in rotamer state r_j. In all calculations, $m = 100$ was used.

All α values in expressions (20) can be determined prior to running the SCMFT protocol, since they are fixed. For each dot in the protein, α values exist for burial by the backbone and for burial by each of the sidechains in each of their rotamer states. However, not all these α values are required to compute expressions (20), due to two major simplifications. The first consists in discarding (1) all dots on backbone atoms that are buried by the backbone or by a particular sidechain i in all of its rotamer states r_i, and (2) all dots on atoms of a sidechain i in a given rotamer state r_i that are buried by the sidechain itself, by the backbone or by any other sidechain $j \neq i$ in all of its rotamer states r_j. In fact, all terms in the summations for $\langle \mathrm{SAS}_a \rangle$ and $\langle \mathrm{SAS}_a \rangle_{r_i}$ corresponding to such dots are equal to zero and, therefore, do not contribute to the sum. The same is true for all terms in $\langle {}^*\mathrm{SIS}_a^i \rangle_{r_i}$ and $\langle {}^*\mathrm{SIS}_a^i \rangle_{r_i, r_j}$ for dots buried by the backbone or by a particular sidechain $k \neq i$ in all of its rotamer states. However, the terms corresponding to dots buried by sidechain i in all of its rotamer states, are not equal to zero. Nevertheless, for a particular dot, the term is constant for all rotamer states of sidechain i and, therefore, contributes a constant value to $\langle \Delta G_i^{\mathrm{des}} \rangle_{r_i}$ for

all r_i, which factors out of the numerator and the denominator of expression (19) and cancels. Under this first simplification, the summations over s_a in expressions (20) reduce to summations over only those dots that are retained. For these dots, all α factors preceding the product of summations in the integrands in expressions (20) equal one. Therefore, these integrands reduce to the respective product of summations for $\langle \mathrm{SAS}_a \rangle$ and $\langle \mathrm{SAS}_a \rangle_{r_i}$, and to the $1 - \alpha^i$ factor times the respective product of summations for $\langle {}^*\mathrm{SIS}_a^i \rangle_{r_i}$ and $\langle {}^*\mathrm{SIS}_a^i \rangle_{r_i, r_j}$.

The second simplification consists in discarding from the products of summations all summations that correspond to sidechains that do not bury the dot in question in any rotamer state. In fact, the summation corresponding to such a sidechain is equal to one and, therefore, does not contribute to the product. Under this simplification, the products reduce to products over only those sidechains that do bury the dot in at least one rotamer state. The consequence of this simplification is that, for each of the dots that are retained from the first simplification, only those rotamers that *do* bury the dot need be determined and stored. The summations over the sidechains to which these rotamers belong can be calculated as:

$$\sum_{r_i} p_i(r_i)\alpha^i(\mathbf{s}_a, r_i) = 1 - \sum_{r_i} p_i(r_i)\Big[\, 1 - \alpha^i(\mathbf{s}_a, r_i)\Big] \tag{21}$$

for a particular sidechain i. The summation on the right-hand side of this expression is simply the sum of the probabilities of sidechain i for the rotamer states in which it *does* bury dot s_a.

The two simplifications described above lead to an enormous enhancement in the computational speed of expressions (20). As a result, introduction of solvation through the methodology developed here is extremely computationally efficient. For the 11 proteins on which sidechain prediction was carried out (see below), solvation led only to an average increase of 7.6% in the total computation time. Moreover, the maximum increase was only 16.5%. The minimum increase was 1.4% .

Derivation of atomic solvation parameter sets

Six sets of atomic solvation parameters were tested in this work. These were derived from three sets of solvation free energies of amino acid sidechain analogs, obtained from the solvation free energies reported by Wolfenden and co-workers [40]. The three sets of energies correspond to three approaches suggested to extract microscopic parameters from macroscopic vapour/water partition experiments [4, 41, 42]. The first set of free energies was the molarity-based solvation free energies, ΔG_ρ, which are proportional to the ratio of the equilibrium molarities of the solute in the vapour and the solvent phases, and correspond directly to the free energies of Wolfenden and co-workers; these free energies correspond to the approach suggested by Ben-Naim [41]. The second set of free energies was the mole-fraction-based solvation free energies, ΔG_x, which are proportional to the ratio of the equilibrium molar fractions of the solute in the vapour and the solvent phases, and that were calculated as:

$$\Delta G_x = \Delta G_\rho - RT \ \ln \frac{p\ V^\circ_{\mathrm{H_2O}}}{RT} \tag{22}$$

where p is the pressure and $V^\circ_{\mathrm{H_2O}}$ is the molar volume of pure liquid water; a value of 1.0 atm was used for p and of 298 K for T. This expression assumes an infinitely dilute solution and ideal gas behaviour of the solute in the vapour phase. These free energies are suggested from the comparative analysis of the solution of a given solute in different globular solvents [4]. The third set of free energies was the Flory-Huggins-type correction to ΔG_ρ described by Sharp and co-workers [42], ΔG_{FH}.

For each set of solvation free energies, two atom-type decompostions of the solute molecule were considered. The first decomposition contained five atom types, which were the same as those previously defined by Eisenberg and McLachlan [11] and by Wesson and Eisenberg [12]: carbon (C), neutral nitrogen or oxygen (N/O), charged nitrogen (N^+), charged oxygen (O^-) and sulphur (S). The second decomposition contained seven atom types: apolar aliphatic carbon (C_{ala}), polar aliphatic carbon (C_{alp}), apolar aromatic carbon (C_{ara}), polar aromatic carbon (C_{arp}), neutral nitrogen or oxygen (N/O), charged nitrogen or oxygen (N^+/O^-) and sulphur (S), where a carbon atom was considered polar if it was bonded to at least one polar or charged N, O or S atom and apolar otherwise. The rationale behind this decomposition is that, firstly, the solvation properties of aromatic molecules are very different to those of aliphatic molecules and, secondly, a carbon atom bonded to a polar or charged N, O or S atom will acquire a partial charge by the inductive effect and, thus, should have solvation properties that are more similar to those of a polar than of an apolar atom. For both atom-type decompositions, both oxy-

gen atoms of Asp and Glu and all three nitrogen atoms of Arg were considered charged.

Optimal values for the atomic solvation parameters of each of the six sets were derived by fitting the calculated solvation free energies to the experimental energies using multiple linear regression (Table 1), where calculated energies were obtained with the ASP model (Equation 2) and the average solvent accessible surfaces reported by Wesson and Eisenberg [12]. To maintain consistency, the solvated atomic radii used to calculate Δs for expressions (20) were those used by Wesson and Eisenberg to calculate the solvent accessible surfaces of the model compounds.

Test systems

The MF methodology described above was applied to the prediction of sidechain conformation and to the prediction of free energies of folding of protein mutants.

Prediction of sidechain conformation

For prediction of sidechain conformation, high quality NMR solution structures of a set of 11 proteins was used as the test set. The PDB codes of these proteins were: 1a2i, 1bmx, 1gb1, 1hun, 1mdj, 1nor, 1pfl, 1pis, 1wjd, 2cbh and 2ezz. The choice of solution structures over crystal structures ensures complete consistency with the energy model used in the methodology, which only takes into account intramolecular interactions and solvation, and therefore is only strictly applicable to a protein molecule in solution isolated from all other protein molecules. We expected solvation to have a more pronounced effect on the prediction of the conformation of surface sidechains, because it makes a larger contribution to the total energy of these sidechains than to that of buried sidechains. Therefore, an analysis of the prediction of surface sidechains would be important. However, in NMR structures, surface sidechains are frequently very disordered. Analysing the prediction of such disordered sidechains has no discriminatory value, since we found that the predicted conformation almost always corresponds to at least one of the possible conformations found in the various NMR models. Therefore, the criterion for the selection of the 11 proteins was that they had a high proportion ($> 65\%$) of well-defined surface residues. A residue was considered well-defined if (1) the various conformations of its sidechain in the set of NMR models of the protein structure formed, at most, two clusters, and (2) within each cluster the standard deviations of the sidechain χ_1 and χ_2 torsion angles were simultaneously less than 20°. This criterion excludes sidechains that are disordered, but does not exclude sidechains that are ordered yet have a certain degree of conformational freedom, as is expected for surface residues. A residue was considered a surface residue if its average relative solvent accessibility over the various NMR models was greater than 30%; relative solvent accessibility of a residue Xaa was defined as the ratio of its solvent accessible surface area in the protein and its solvent accessible surface area in the extended Gly-Xaa-Gly tripeptide. Relative solvent accessibilities were calculated with the WHATIF package [43].

Backbone and prosthetic group atomic co-ordinates for the MF calculations were taken from the NMR model that was closest to the average structure calculated from all models. The sidechain conformations of all residues were predicted for each protein. Predicted models for the sidechains were built as described in [21]. Sidechain prediction accuracy was evaluated as the fraction of well-defined residues for which the χ_1 and χ_2 torsion angles of the predicted sidechain conformation were simultaneously within 40° (or 20° for Pro) of the corresponding experimental sidechain conformation in any one of the various NMR models [34].

Prediction of folding free energies

For the prediction of folding free energies of mutant proteins, we used a test set corresponding to all apolar and polar single mutants of barnase published in an extensive study performed in the Fersht group [24-30], and compiled in [26]. A mutant was termed 'apolar' if the mutated residue was apolar in both the wild type (WT) protein and the mutant protein, and 'polar' if it was polar in either protein. For all proteins, the free energy of the native state was calculated using the backbone co-ordinates of the WT protein (PDB code 1a2p), and the free energy of the denatured state was calculated using a single extended backbone conformation in which φ was equal to $-122°$ and ψ was equal to 143°; these values correspond to the potential energy minimum in the β-sheet region for the acteyl-alanyl-aminomethyl dipeptide calculated with the vacuum GROMOS force-field [35]. In all calculations, the sidechain conformations of all 98 non-Gly residues were predicted, for both the native and the denatured state. Free energies were calculated with expression (3) for the solvated system and the corresponding expression defined in [32] for the non-solvated system.

Table 1. Atomic solvation free energy parameters and respective standard deviations in cal mol^{-1} Å^{-2}. Details of the derivation of the parameters are given in the text.

		ΔG_ρ	ΔG_x	ΔG_{FH}
5-atom-type	C	−0.16 ± 3.32	17.7 ± 4.3	11.3 ± 3.3
decomposition	N/O	−120.4 ± 14.2	−96.0 ± 18.6	−118.3 ± 14.3
	N^+	−161.1 ± 15.3	−148.0 ± 20.0	−154.7 ± 15.4
	O^-	−146.8 ± 20.1	−118.9 ± 26.3	−145.0 ± 20.2
	S	−19.7 ± 23.0	11.3 ± 30.1	−16.6 ± 23.1
7-atom-type	C_{ala}	9.6 ± 1.5	30.2 ± 3.0	21.1 ± 1.2
decomposition	C_{alp}	−11.0 ± 5.0	14.5 ± 10.4	−3.6 ± 4.0
	C_{ara}	−9.3 ± 2.3	3.4 ± 4.7	2.9 ± 1.8
	C_{arp}	−48.7 ± 5.9	−36.0 ± 12.3	−39.2 ± 4.8
	N/O	−110.2 ± 4.9	−88.5 ± 10.2	−106.2 ± 4.0
	N^+/O^-	−163.7 ± 4.4	−151.8 ± 9.2	−157.3 ± 3.6
	S	−6.9 ± 10.2	11.5 ± 21.3	2.5 ± 8.2

Folding free energies were calculated as the difference between the free energies of the native and the denatured states.

The results of the folding free energy predictions were analysed using two different approaches. In the first, the correlation between calculated and experimentally observed folding free energy differences of the mutant proteins relative to the WT protein was evaluated. The statistical significance of the correlations was analysed with a *t*-Student test [44] at 5% level of significance. In the second approach, we assessed the accuracy with which the qualitative effect on protein stability – stabilisation or destabilisation – produced by the mutations can be predicted based on the calculated folding free energies. This was performed as follows: all possible $\frac{1}{2}N\ (N-1)$ pairs of protein mutants were formed, where N is the total number of mutants plus WT. This corresponds to 231 apolar mutant pairs ($N = 22$) and 351 polar mutant pairs ($N = 27$). These pairs were then grouped into experimental folding free energy classes of 0.5 kcal mol^{-1} in width. For the apolar mutants, the 0.0–0.5, 0.5–1.0, 1.0–1.5, 1.5–2.0, 2.0–2.5, 2.5–3.0, 3.0–3.5, 3.5–4.0, 4.0–4.5 and 4.5–5.0 kcal mol^{-1} classes contained respectively 67, 48, 18, 12, 20, 22, 21, 17, 4 and 2 mutant pairs. For the polar mutants, the 0.0–0.5, 0.5–1.0, 1.0–1.5, 1.5–2.0, 2.0–2.5, 2.5–3.0 and 3.0–3.5 kcal mol^{-1} classes contained respectively 99, 87, 75, 59, 35, 9 and 14 mutant pairs. The prediction accuracy was then calculated within each class as the fraction of mutant pairs for which stabilisation or destabilisation produced by the mutation was correctly predicted from the calculated folding free energies of the two proteins in the pair. For each class it was analysed whether the prediction accuracy obtained was significantly higher than random (50% accuracy) using a binomial test [44] at 5% level of significance.

Results and discussion

Prediction of sidechain conformation

Introduction of solvation into the SCMFT method resulted in an increase in overall sidechain prediction accuracy for all three of the five-atom-type ASP sets (Figure 1 All). For the ΔG_ρ set, the increase was only marginal, but for the other two sets, although still small, the increase was significant. This increase in overall prediction accuracy was due to a large increase in the prediction accuracy of charged sidechains for all three ASP sets, and to a small increase in that of apolar sidechains for the ΔG_x and ΔG_{FH} sets (Figure 1 All). In contrast, the prediction accuracy of polar sidechains decreased for all three sets.

A separate analysis of buried sidechains and surface sidechains revealed a small decrease in the overall prediction accuracy of buried sidechains with the introduction of solvation, for all three of the five-atom-type ASP sets (Figure 1 Buried). For each of these sets, the prediction accuracy of charged, polar and apolar sidechains decreased. However, the decrease was large only for charged sidechains and, for the ΔG_x set in

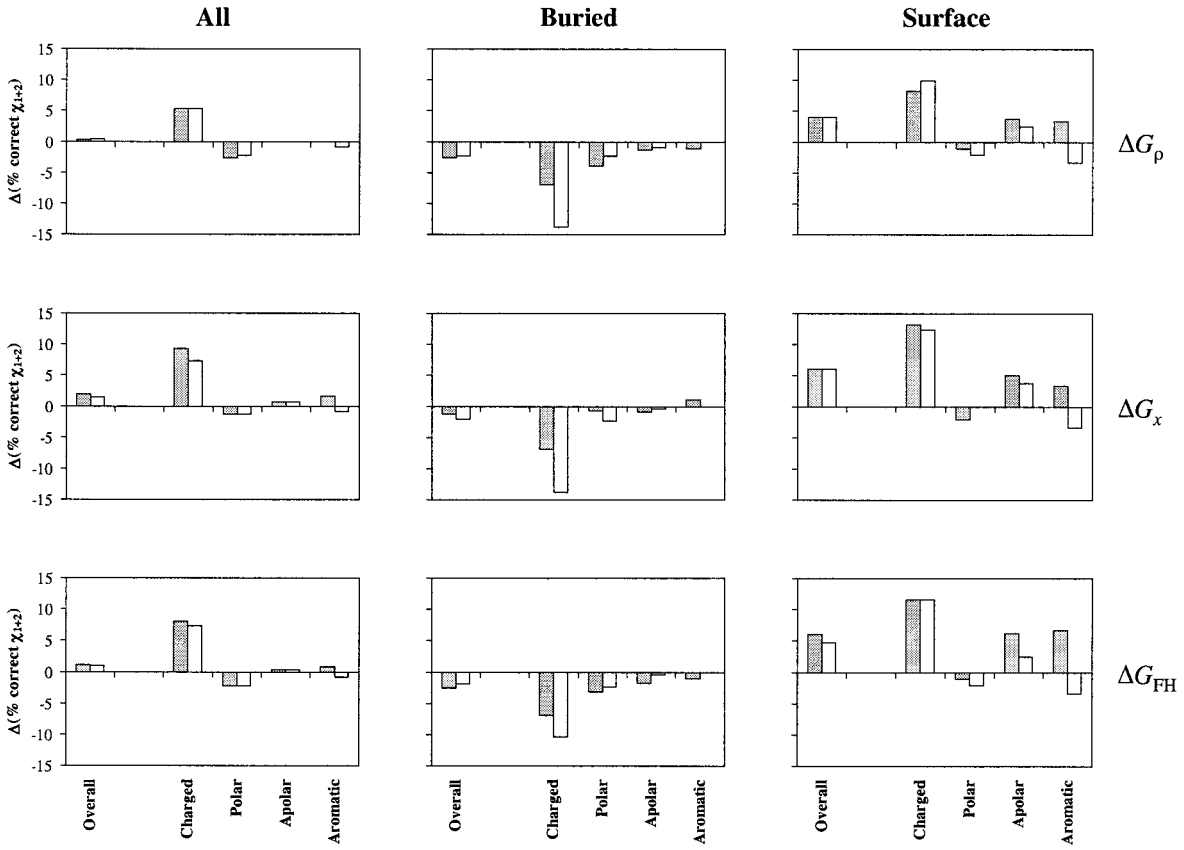

Figure 1. Effect of solvation on prediction accuracy of well-defined sidechains: variation in the χ_{1+2} prediction accuracy resulting from inclusion of solvation in the SCMFT method, relative to the predictions obtained without solvation. Results are given for all sidechains considered together, and for buried and surface sidechains considered separately. The charts on each line represent the results obtained with the ASP sets derived from the model compound solvation free energies indicated to the right of the line (ΔG_ρ – molarity-based solvation free energies, ΔG_x – mole-fraction-based solvation free energies, ΔG_{FH} – solvation free energies obtained by a Flory–Huggins-type correction to ΔG_ρ; see Materials and Methods for details on the derivation of the ASP sets.). Grey bars: five-atom-type ASP set. White bars: seven-atom-type ASP set. Residue type classes are defined as: Overall – all residue types; Charged – Asp, Glu, Arg, Lys; Polar – Ser, Thr, Asn, Gln, His, Tyr, Trp; Apolar – Cys, Met, Pro, Val, Leu, Ile, Phe; Aromatic – His, Phe, Tyr, Trp. Note that Overall = Charged + Polar + Apolar; Aromatic is a subclass of Polar + Apolar. The absolute prediction accuracies obtained without solvation were: All – Overall 67.5%; Charged 48.0%; Polar 66.5%; Apolar 77.9%; Aromatic 81.3% . Buried – Overall 79.2%; Charged 65.5%; Polar 71.1%; Apolar 85.5%; Aromatic 86.0%. Surface – Overall 52.5%; Charged 43.8%; Polar 60.4%; Apolar 56.3%; Aromatic 66.7%.

particular, it was only marginal for polar and apolar sidechains.

In contrast to the results for buried sidechains, solvation resulted in a substantial increase in the overall prediction accuracy of surface sidechains, for all three of the five-atom-type ASP sets (Figure 1 Surface). This increase was due to a large increase in prediction accuracy for charged sidechains and to a smaller, but still substantial, increase for apolar sidechains. On the contrary, polar sidechains suffered a small decrease in prediction accuracy for all three ASP sets. With the exception of the polar sidechain class, the absolute variations in prediction accuracy resulting from solvation were larger for surface sidechains than for buried sidechains, for overall and for each of the individual sidechain classes. This is what would be expected from the fact that solvation free energy corresponds to a larger fraction of the total energy of surface sidechains than of buried sidechains and, therefore, is more important in determining the conformation of surface sidechains.

An important conclusion can be drawn from the results obtained for surface sidechains as follows. The conformation of charged surface-sidechains is determined predominantly by the tendency to maximise exposure to solvent, and that of apolar surface-sidechains predominantly by the tendency to minimise this exposure. Therefore, the increase in prediction accuracy

obtained for charged and apolar surface sidechains suggests that the combined ASP/distance dependent dielectric constant model accounts for these general details of solvation relatively well. The conformation of polar surface-sidechains, on the other hand, is determined by a more subtle balance between solvation and the interaction of the sidechain with other polar or charged groups in the protein. Therefore, the small decrease in prediction accuracy obtained for polar surface sidechains indicates that this solvation model does not capture these fine details of solvation, at least not with the ASP sets tested here.

The overall sidechain prediction accuracy obtained with each of the seven-atom-type ASP sets was only marginally different to that obtained with the corresponding five-atom-type set (Figure 1 All). For the individual charged, polar and apolar sidechain classes, the difference in prediction accuracy was also not significant, except for the charged sidechain class for the ΔG_x set, where prediction accuracy with the seven-atom-type set was substantially lower. For buried sidechains, the overall prediction accuracies obtained with the two decompositions were also only marginally different for all three pairs of ASP sets (Figure 1 Buried). The same occurred with surface sidechains, except for the ΔG_{FH} sets, for which the prediction accuracy of the seven-atom-type set was slightly lower (Figure 1 Surface). For the individual charged, polar and apolar sidechain classes, the only consistent variations in prediction accuracy across the three pairs of ASP sets were for charged and apolar buried-sidechains and for apolar surface-sidechains: for apolar buried-sidechains, prediction accuracy was consistently higher with the seven-atom-type sets, but both for charged buried-sidechains and for apolar surface-sidechains, it was consistently lower for these sets (Figure 1 Buried and Surface). For aromatic surface-sidechains – a subclass of the polar and apolar classes – prediction accuracy was consistently considerably lower for the seven-atom-type sets (Figure 1 Surface).

The results from this comparison were unexpected. In fact, the seven-atom-type decomposition is in principle physically more detailed than the five-atom-type decomposition. Correspondingly, it led to a considerably better fit to the model compound data, portrayed by the lower standard deviations of the parameters (see Table 1). Therefore, we had anticipated obtaining substantially better predictions with the seven-atom-type ASP sets. Particularly unexpected was the consistently lower prediction accuracy of aromatic surface-sidechains with these sets, since aromatic and aliphatic carbons were treated separately in the seven-atom-type decomposition but jointly in the five-atom-type decomposition. We believe that these results may be due to the errors in solvation energy inherent to the use of the rigid rotamer model (RRM) (for a detailed discussion on this topic see [32] and [34]). In fact, we obtained similar results in another study [32] using the RRM where, above a certain level of physical detail, an increase in the physical detail of the potential energy function did not lead to an expected increase in sidechain prediction accuracy.

From a fundamental point of view, the results obtained here show that solvation does increase the sidechain prediction accuracy of the SCMFT method. From a practical point of view, even though the increase in overall prediction accuracy obtained for All sidechains is small, introducing solvation is still worth the additional computational cost. First, because including solvation results only in a small 7.6% average increase in computation time. Second, because although the increase in prediction for All sidechains is small, the increase for surface sidechains is substantial. Surface sidechains play a central role in determining many of the protein's functions and/or properties, such as its interaction with ligands and with other proteins, and its overall pK_a. Therefore, the degree of increase in prediction of surface sidechains obtained here may allow a more accurate prediction of these functions and/or properties by other methods.

Of the six ASP sets studied here, the five-atom-type ΔG_x set is the best for sidechain prediction. In fact, not only does this set lead to the highest increase in overall prediction accuracy of All sidechains, but it results in a large increase in the prediction accuracy of charged and apolar surface-sidechains and only a marginal decrease in that of polar and apolar buried-sidechains.

Prediction of folding free energies of barnase mutants

Apolar mutants

For the apolar barnase mutants, the calculated variations in folding free energy relative to the WT protein correlated with the corresponding experimentally observed quantities, both for the calculations with solvation and for those without (Figure 2). In all cases, the correlation was statistically significant in a *t*-Student test at 5% level of significance. The general trend of the data (represented by the positive slope of the best fit line) was correct, that is, an increase in the ex-

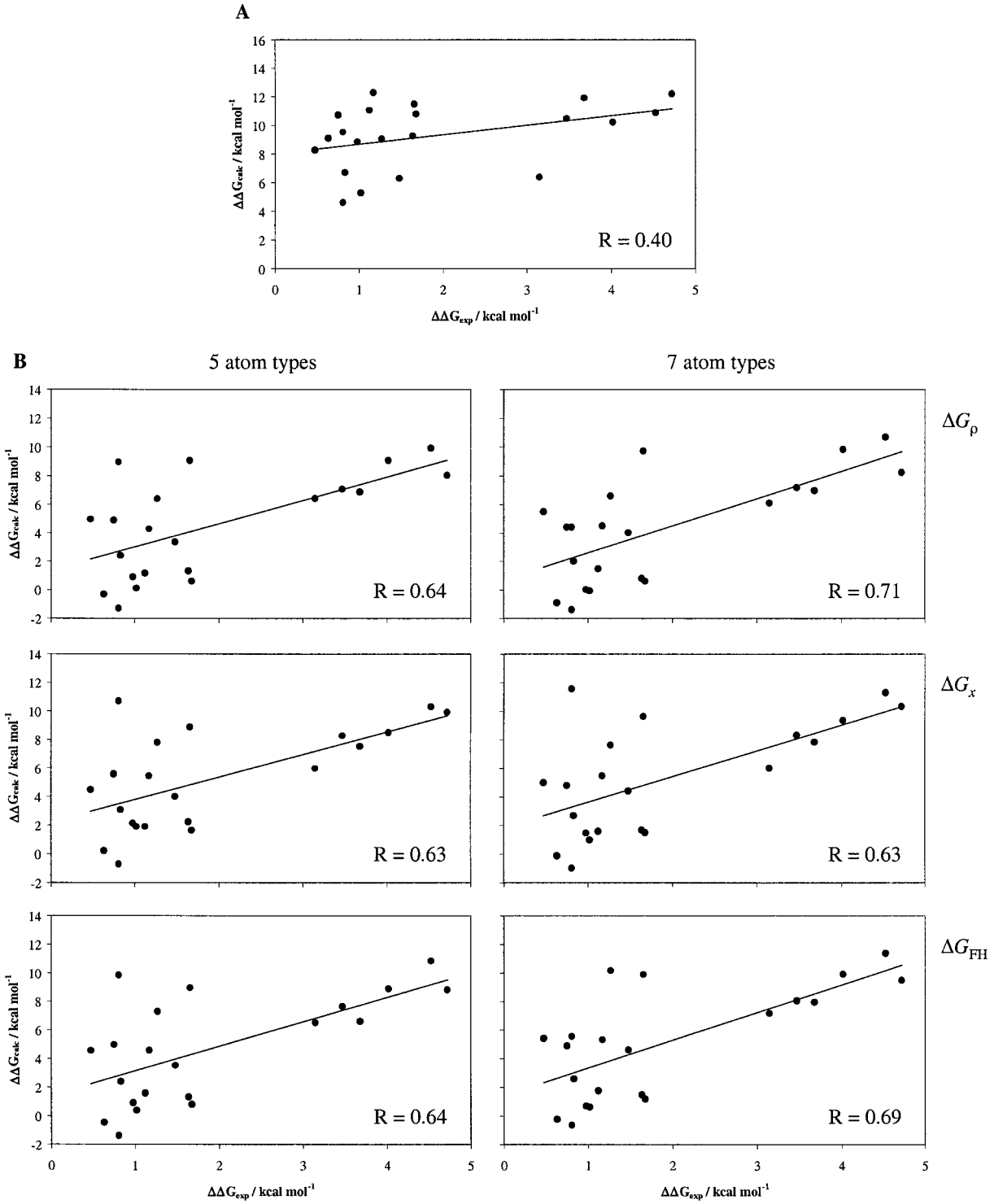

Figure 2. Effect of solvation on the prediction of folding free energies of apolar barnase mutants: calculated versus experimentally observed variation in folding free energy relative to WT. A. calculations performed without solvation. B. calculations performed with solvation. Charts on the same line in B correspond to ASP sets derived from the same solvation free energies, indicated to the right of each line. Charts in the same column in B correspond to the same atom-type ASP decomposition, indicated at the top of each column. The correlation coefficient R and the best fit line are shown in each chart.

perimentally observed folding free energy of a given mutant in relation to another corresponded, in general, to an increase in the respective calculated quantity. The range spanned by the calculated folding free energy variations (about 8 kcal mol^{-1} for calculations without solvation and 12 kcal mol^{-1} for calculations with solvation) was on the same order of magnitude as the range spanned by the observed folding free energy variations (4.25 kcal mol^{-1}). The results obtained here show that the MF free energy function involving a rotamer description of sidechain conformational space is sufficiently accurate to be used for prediction of folding free energies of mutants, as we had set out to investigate.

Concerning the effect of solvation described by an ASP potential on the correlation between calculated and observed folding free energy variations, we found that introduction of solvation yielded in a substantial improvement in correlation for all six ASP sets in relation to the results obtained without solvation (Figure 2). This result is in agreement with the fact that solvation is important in determining the folding free energy of a protein [45]. The improvement in correlation was essentially the same for all three of the five-atom-type ASP sets. Correlation improved with the seven-atom-type decomposition in relation to the five-atom-type decomposition for the ΔG_ρ and ΔG_{FH} ASP sets, but was the same with both decompositions for the ΔG_x sets. These results are as would be expected from the, in principle, more physical seven-atom-type decomposition. The best correlation overall was obtained with the seven-atom-type ΔG_ρ ASP set, which is also in opposition to the results obtained for sidechain prediction.

The correlations observed in Figure 2 are due primarily to the existence of correlation for mutants with a large difference between their experimental folding free energies, since essentially no correlation exists for mutants for which this energy difference is small. In fact, if regression is performed separately for the subset of mutants with $\Delta\Delta G_{\text{exp}} < 2$ kcal mol^{-1} and for the subset with $\Delta\Delta G_{\text{exp}} > 3$ kcal mol^{-1}, the correlations obtained for the first subset are not statistically significant in a *t*-Student test at 5% significance level, either for the calculations without solvation ($R = 0.292$) or for any of the calculations with solvation ($R <= 0.231$). For the other subset, correlation is statistically significant at 5% significance for all calculations with solvation ($R => 0.750$), in spite of the differences between the experimental folding free energies for the six mutants in this subset being small. We were unable to find any structural-based explanation for the different behaviour of the mutants in the two subsets. It is possible that the correlation observed for the higher energy subset is due to the small number of points not representing a true random sample, whereby application of the statistical test may not be valid. As such, we prefer to be conservative in our analysis of these results, and consider that the methodology described here is, in general, not accurate enough to predict small folding free energy differences, but only trends for large energy differences. This will be analysed more quantitatively below in the subsection 'Predicting stabilisation and destabilisation'.

It is interesting to note that one of the sources of correlation between calculated and observed folding free energy variations was the model of the denatured state used in the present methodology. In fact, there is now a large amount of experimental and theoretical evidence that the denatured state of a protein under physiological conditions is an ensemble of relatively compact conformations with considerable secondary structure, and that mutations may act not only on the stability of the native state of the protein but also on that of the denatured state through short-range sidechain-sidechain interactions and variations in solvation free energy, which arise from the close spatial proximity of sidechains in secondary structure elements [46, 47]. Such interactions and their effect on folding free energies are completely neglected in commonly-used random coil models of the denatured state, which assume that there are no sidechain-sidechain interactions, sidechains are completely exposed to solvent and sidechain conformational entropy is maximum [47, 48]. When we tested such a model [49], in which solvation free energies of the completely exposed sidechains were calculated using the ASP sets derived in the present work and maximum conformational entropy was calculated from the number of rotamers in the rotamer library (results not shown), we found a systematic decrease of 0.1 to 0.2 in correlation for all calculations with solvation. These results suggest that, although the single extended conformation model of the denatured state used in the present work models neither the conformational multiplicity of the denatured state nor the compactness of the conformations comprised in this state, it does apparently capture some of the general features of this state. In particular, we observed that sidechain conformational entropy was very much lower than the maximum entropy for some residues.

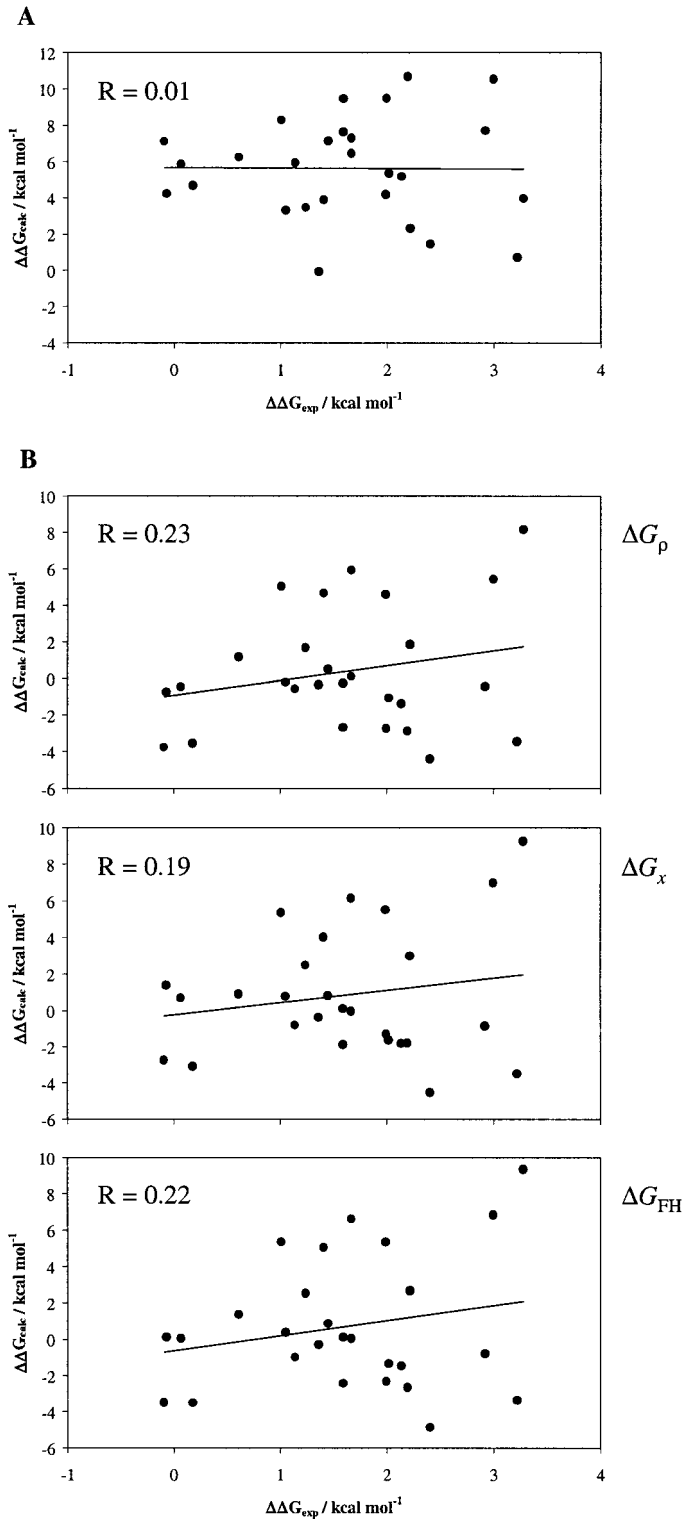

Figure 3. Effect of solvation on the prediction of folding free energies of polar barnase mutants: calculated versus experimentally observed variation in folding free energy relative to WT. A. calculations performed without solvation. B. calculations performed with solvation. All calculations with solvation were performed with the seven-atom-type ASP decomposition, based on the better results obtained for this decomposition with apolar mutants. The solvation free energies from which the ASP set corresponding to each chart in B was derived are indicated to the right of the chart. The correlation coefficient R and the best fit line are shown in each chart.

Polar mutants

For the set of polar barnase mutants, we found that the calculated and experimentally observed variations in folding free energy relative to WT were totally uncorrelated, for the calculations without solvation (Figure 3). Introduction of solvation led to a significant improvement in correlation, but it was still not significant in a *t*-Student test at 5% level of significance. As occurred with the apolar mutants, the highest correlation was obtained with the ΔG_ρ ASP set.

The very low correlation obtained for polar mutants suggests that the solvation free energies obtained for polar residues with the combined ASP/distance dependent dielectric constant model are too crude for prediction of folding free energies of protein mutants that differ in the number of these residues. In fact, although polar residues do also exist in the apolar mutants studied in the present work, all mutants had the same set of polar residues. Therefore, the sum of the errors in the solvation free energy of polar residues would be approximately the same in all mutants and, thus, would in part cancel out in the folding free energy variations, whence the observed correlation. On the contrary, the polar mutants studied in the present work differed by one or two polar residues. In this case, the sum of the errors in the solvation free energy of polar residues would not in general be the same in all mutants and, thus, could not cancel out in the folding free energy variations; this may be the source of the very low correlation observed.

Finally, we point out that the significant improvement in correlation observed for the apolar mutants after introduction of solvation is consistent with the increase in sidechain prediction accuracy observed for apolar sidechains in the previous section. Likewise, the lack of correlation for polar mutants even after introducing solvation agrees with the small decrease in sidechain prediction accuracy seen for polar sidechains in the previous section.

Predicting stabilisation or destabilisation

From the correlations presented in the previous two subsections, it was concluded that at its present stage of development, the methodology described here is not accurate enough to predict small experimental folding free energy differences, but it can allow the trends (*i.e.* increase or decrease) of the folding free energies to be predicted when the experimental free energy differences are large. However, from a practical point of view such trends are still useful, since in a mutation study it is frequently sufficient to know if a given mutation will stabilise or destabilise the protein. The results obtained for the apolar and polar barnase mutants can be analysed from this perspective by forming all possible pairs of mutants, considering one protein in a pair as the WT protein and the other as the mutant, and assessing the accuracy with which the qualitative effect of the mutations on protein stability (stabilisation or destabilisation) can be predicted using the calculated folding free energies. This analysis also allows a quantitative assessment of the conclusion stated above.

For apolar mutant pairs, the prediction accuracy obtained without solvation was significantly higher than random for the three lowest energy classes, and increased progressively as $\Delta\Delta G_{exp}$ increased (Figure 4). However, for the next two classes (1.5–2.0 and 2.0–2.5 kcal mol^{-1}) prediction decreased to values not significantly higher than random. Only from the 2.0–2.5 kcal mol^{-1} class onwards was there a consistent increase in prediction accuracy with $\Delta\Delta G_{\rm exp}$. From a physical perspective, it is reasonable to expect that the larger the experimental folding free energy difference between the mutants, the easier it should be to correctly predict stabilisation or destabilisation. Therefore, this result is as expected. The two highest energy classes require some discussion. Although the prediction accuracy is at 100% , it is not significantly higher than random in a binomial test at 5% significance level. However, this is due to the number of mutants in each class being too small to allow the statistical significance of the results to be evaluated at this level of significance. Nevertheless, they probably do represent the true trend of the data, given that they continue the physically reasonable trend of the previous four energy classes.

Introduction of solvation resulted in an overall improvement in prediction accuracy profile for the apolar mutants (Figure 4). Thus, prediction accuracy was significantly higher than random from the 0.5–1.0 kcal mol^{-1} class onwards for the ΔG_x and $\Delta G_{\rm FH}$ ASP sets and from the 1.0-1.5 kcal mol^{-1} class onwards for the ΔG_ρ ASP set. Moreover, prediction accuracy began to increase consistently from the lowest 0.0-0.5 kcal mol^{-1} class, compared to the 2.0-2.5 kcal mol^{-1} class seen above for the calculations without solvation.

The data can also be analysed from a slightly different perspective, by determining the value of $\Delta\Delta G_{\rm exp}$ above which a high prediction accuracy is achieved. In fact, in a mutation study, only high

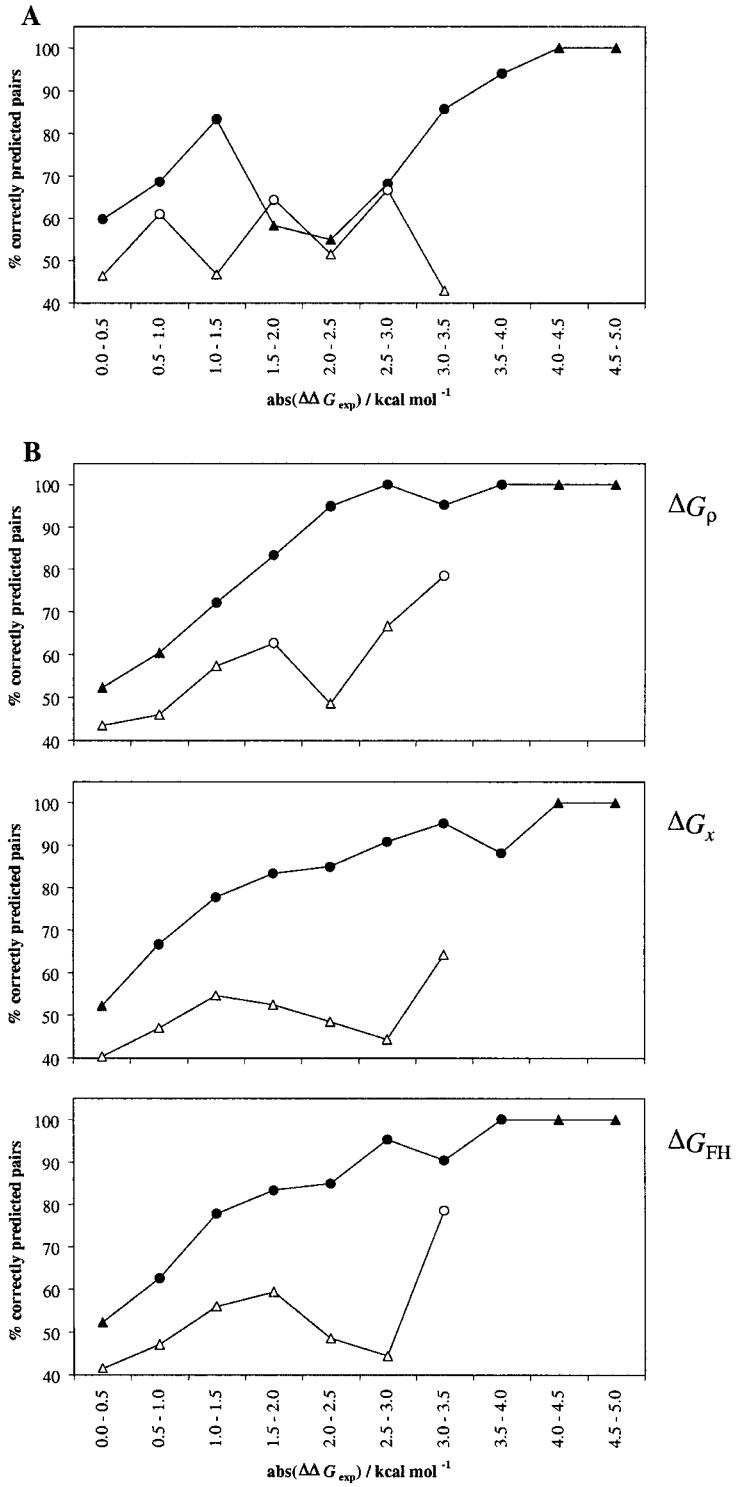

Figure 4. Prediction of the qualitative effect of mutation on protein stability (stabilisation or destabilisation) as a function of the absolute variation in observed folding free energy resulting from mutation. All possible pairs of barnase mutants were generated and grouped into the free energy classes given on the horizontal axis. Prediction accuracy within each class was measured as the fraction of mutant pairs for which stabilisation or destabilisation resulting from mutation was correctly predicted (see Materials and Methods for details). A. Calculations performed without solvation. B. Calculations performed with solvation. All calculations with solvation were performed with the seven-atom-type ASP decomposition. The solvation free energies from which the ASP set corresponding to each chart in B was derived are indicated to the right of the chart. Black symbols: apolar mutants. White symbols: polar mutants. Circles: prediction accuracy significantly higher than random prediction (50%) in a binomial test at 5% level of significance. Triangles: prediction accuracy not significantly higher than random at 5% level of significance.

prediction accuracy is of any practical utility. For the calculations without solvation, a high prediction accuracy (> 90%) was only achieved for mutant pairs with folding free energy differences in the 3.5-4.0 kcal mol^{-1} range (Figure 4). Introduction of solvation improved this value. Thus, for the ΔG_ρ ASP set, a high prediction accuracy was achieved for mutant pairs with folding free energy differences in the 2.0–2.5 kcal mol^{-1} range, which is considerably lower that that obtained without solvation. The results obtained with the ΔG_x and ΔG_{FH} ASP sets were slightly inferior, a high prediction accuracy only having been attained for mutant pairs with folding free energy differences in the 2.5-3.0 kcal mol^{-1} range. In conclusion, the significance of the results obtained here to a mutation study is that, if mutations are restricted to apolar residue types, the effect of mutation on protein stability should be essentially correctly predicted if the absolute variation in folding free energy arising from the mutations is higher than 2.0–2.5 kcal mol^{-1}, when the best set of conditions is used (solvation using the ΔG_ρ ASP set).

For polar mutant pairs, the prediction accuracy was, in general, lower than that obtained for apolar mutants both without and with solvation (Figure 4). Without solvation, although prediction accuracy was significantly higher than random for three energy classes, it increased and decreased randomly as the difference in folding free energies increased and, therefore these predictions are of no practical value. This was consistent with the complete lack of correlation between calculated and observed folding free energies. With solvation, the prediction accuracy obtained with the ΔG_x ASP set was not significantly higher than random for any of the energy classes. For the ΔG_ρ ASP set, there appeared to be a consistent increase in prediction accuracy that began at the 2.0-2.5 kcal mol^{-1} class. However, only the prediction accuracy of the highest of these three classes (3.0–3.5 kcal mol^{-1}) was statistically significant, and a high prediction accuracy had not yet been achieved for this class. If the observed trend in the data were maintained, a high prediction accuracy would only be attained for folding free energy differences in the 4.0-4.5 kcal mol^{-1} range. For the ΔG_{FH} ASP set, there was a large increase in prediction accuracy between the two classes of highest free energy differences. A consistent increase may have been seen for higher free energy differences, but no higher observed differences existed in the set of mutant pairs studied in this work. In conclusion, the significance of the results obtained here to a mutation study is that, if mutations involve both apolar and polar residue types, the effect of mutation on protein stability might only be essentially correctly predicted if the absolute variation in folding free energy arising from the mutations is higher than 4.0-4.5 kcal mol^{-1}, when the best set of conditions is used (solvation using the ΔG_ρ ASP set).

Future developments in methodology

Several improvements in the methodology described in the present work may be introduced in the future. First, although the average intramolecular potential energies were calculated with the FRM, the average solvation free energies were calculated with the classic RRM. An extension of the FRM to calculation of solvation free energies is not straightforward when solvation free energy is described by the ASP model, due to the non-decomposability of SAS into pairwise terms, and the theoretical framework for such an extension has not yet been developed. Nevertheless, such a development is likely to improve the accuracy of these free energies and, consequently, both the accuracy with which sidechain conformation is predicted and the correlation between calculated and observed folding free energy variations of protein mutants. Indeed, the average intramolecular potential energies calculated with FRM are substantially more accurate than those calculated with RRM. Thus, in the context of sidechain prediction, prediction accuracy obtained with FRM was considerably higher than that obtained with RRM [34]. Likewise, in the context of folding free energy prediction, we obtained a correlation of 0.11 for the apolar mutants with RRM for the calculations without solvation (results not shown), which is 0.3 lower than the correlation of 0.40 obtained with FRM (see Figure 2). Moreover, the range spanned by the folding free energy variations calculated with RRM (about 22 kcal mol^{-1}) was considerably higher than that spanned by the folding free energy variations calculated with FRM (about 8 kcal mol^{-1}), and five times larger than that spanned by the observed folding free energy variations (4.25 kcal mol^{-1}). These results suggest that the FRM may be a solution to a serious limitation of most current rotamer-based methods for *de novo* protein design – high sensitivity of van der Waals and hydrogen-bonding terms to deviation of the RRM from optimal interaction geometry – which appears to be one of the main sources of their significant false negative rate [48].

Second, a rigid model of the backbone, corresponding to the co-ordinates of WT, was used for the native state in all folding free energy calculations. However, it is now well established that the protein backbone can shift to accommodate mutations [50, 51]. A generalised model for backbone flexibility, that correctly accounts for the direction and magnitude of the backbone shift that accompanies a particular mutation, might improve the accuracy of calculated folding free energy variations. Such a generalised model has not yet been developed, but some efforts in this direction have been made in related areas [52–54].

Third, we found that the combined ASP/distance-dependent dielectric constant solvation model led to a small decrease in the prediction accuracy of polar residues and was too crude to allow prediction of folding free energy variations of polar mutants. Implementation of a more detailed solvation model into the SCMFT method may improve both the prediction of the sidechain conformations of surface residues and the accuracy of calculated folding free energies variations of polar mutants, as well as apolar mutants. The most common of such models are the continuum Poisson-Boltzmann models [5] and the Langevin-dipoles model [6]. Implementation of such models in the SCMFT method is not straightforward and will require the development of an appropriate theoretical framework. Some progress has been made in this direction with an approximate Langevin-dipoles model [33], but implementation of a complete model has not yet been described. In the event of incorporating these more detailed solvation models in SCMFT, a complementary model to calculate the hydrophobic component of solvation (sum of the van der Waals component and the cavity formation component) will still be required, as described in the introduction. A commonly used hydrophobic solvation model is a simple linear function of the total SAS [8–10], to which the methodology developed in the present work is directly applicable. Therefore, this methodology will be necessary even when more detailed models for electrostatic solvation are implemented in SCMFT.

Conclusions and perspectives

In this work, we present a method for incorporating the Atomic Solvation Parameter (ASP) model for solvation into the Self-Consistent Mean Field Theory (SCMFT) method for sidechain modelling. Until now this had not been accomplished. The theory on which the method is based is exact within the MF framework, that is, it does not depend on any approximation of the solvent accessible surface area (SAS) – intrinsically a many-body quantity. Therefore, unlike Dead-End Elimination methods for sidechain modelling, in which implementation of the ASP potential requires an approximate pairwise decomposition of the many-body SAS [22], in SCMFT the SAS can be calculated to arbitrary accuracy. The method is computationally very efficient: only 7.6% slower on average than the method without solvation. This is an advantage over combinatorial search and Monte Carlo methods for sidechain modelling, where implementation of the ASP potential requires the calculation of SAS of an entire protein conformation in each step of the algorithm [14–16], which can become computationally very expensive. Inclusion of solvation into SCMFT through the method described here substantially improved the prediction accuracy of well-defined surface sidechains. The overall improvement in surface plus buried sidechains was marginal, partly due to a small decrease in overall prediction of buried sidechains but also because the number of buried sidechains overwhelms that of surface sidechains in the global average. For apolar barnase mutants, it also significantly improved the correlation between calculated and observed folding free energy variations relative to the wild type protein. The final correlation of coefficient 0.7 may still be too low to quantitatively predict folding free energy differences, but it does already allow the prediction of trends in these energy differences. For polar mutants, correlation was not significant even with solvation. In conclusion, both for sidechain prediction and prediction of folding free energy differences for protein mutants, inclusion of solvation in the SCMFT method represents a step in the right direction. However, further developments in the methodology have to be made to improve the absolute values of sidechain prediction for surface sidechains and to allow the quantitative prediction of folding free energy differences.

The methodology described in this work can easily be extended to *de novo* protein design. Most limitations and required improvements in current *de novo* protein design methods discussed in a recent review [48] – high sensitivity of van der Waals and hydrogen-bonding terms to deviation of the RRM from optimal interaction geometry, requirement of a solvation free energy term, requirement of a sidechain conformational entropy term, requirement of an improved model for the denatured state – have to a greater or

lesser degree been solved in this methodology. Therefore, the perspectives for such an extension seem very promising.

Acknowledgements

We acknowledge financial support from Fundação para a Ciência e a Tecnologia, Portugal, through grants PRAXIS/P/BIO/14314/1998 and FCT 32789/99. J.M. acknowledges PRAXIS XXI fellowship BD/2740/94. AMB acknowledges PRAXIS XXI fellowship BPD/18899/98.

References

1. Smith, P.E. and Pettitt, B.M., J. Phys. Chem., 98 (1994) 9700.
2. Dill, K.A., Biochemistry, 29 (1990) 7133.
3. Leach, A.R. Molecular Modelling: Principles and Applications. Addison Wesley Longman Ltd., Essex, 1996.
4. Chan, H.S. and Dill, K.A., Ann. Rev. Biophys. Biomol. Struct., 26 (1997) 425.
5. Warwicker, J. and Watson, H.C., J. Mol. Biol., 157 (1982) 671.
6. Warshel, A. and Levitt, M., J. Mol. Biol., 103 (1976) 227.
7. Lee, B. and Richards, F.M., J. Mol. Biol., 55 (1971) 379.
8. Herman, R.B., J. Phys. Chem., 76 (1972) 2754.
9. Chothia, C.H., Nature, 248 (1974) 338.
10. Reynolds, J.A., Gilbert, D.B. and Tanford, C., Proc. Natl. Acad. Sci. USA, 71 (1974) 2925.
11. Eisenberg, D. and MacLachlan, A.D., Nature, 319 (1986) 199.
12. Wesson, L. and Eisenberg, D., Protein Sci., 1 (1992) 227.
13. Janardhan, A. and Vajda, S., Protein Sci., 7 (1998) 1772.
14. Wilson, C., Gregoret, L.M. and Agard, D.A., J. Mol. Biol., 229 (1993) 996.
15. Cregut, D., Liautard, J.-P. and Chiche, L., Protein Eng, 7 (1994) 1333.
16. Cardozo, T., Totrov, M. and Abagyan, R., Proteins, 23 (1995) 403.
17. Schiffer, C.A., Caldwell, J.W., Kollman, P.A. and Stroud, R.M., Mol. Simul., 10 (1993) 121.
18. von Freyberg, B., Richmond, T.J. and Braun, W., J. Mol. Biol., 233 (1993) 275.
19. Lee, C., J. Mol. Biol., 236 (1994) 918.
20. Desmet, J., de Maeyer, M., Hazes, B. and Lasters, I., Nature, 356 (1992) 539.
21. Koehl, P. and Delarue, M., J. Mol. Biol., 239 (1994) 249.
22. Street, A.G. and Mayo, S.L., Folding & design, 3 (1998) 253.
23. Voigt, C.A., Gordon, D.B. and Mayo, S.L., J. Mol. Biol., 299 (2000) 789.
24. Serrano, L., Bycroft, M. and Fersht, A.R., J. Mol. Biol., 218 (1991). 465
25. Fersht, A., Matouschek, A. and Serrano, L., J. Mol. Biol., 224 (1992). 771
26. Serrano, L., Kellis Jr., J.T., Cann, P., Matouschek, A. and Fersht, A.R., J. Mol. Biol., 224 (1992) 783.
27. Serrano, L., Matouschek, A. and Fersht, A.R., J. Mol. Biol., 224 (1992) 805.
28. Matouschek, A., Serrano, L. and Fersht, A.R., J. Mol. Biol., 224 (1992) 819.
29. Matouschek, A., Serrano, L., Meiering, E.M., Bycroft, M. and Fersht, A.R., J. Mol. Biol., 224 (1992) 837.
30. Serrano, L., Matouschek, A. and Fersht, A.R., J. Mol. Biol., 224 (1992) 847.
31. Hill, T.L. Statistical Mechanics, McGraw-Hill, New York, 1956.
32. Mendes, J., Soares, C.M. and Carrondo, M.A., Biopolymers, 50 (1999) 111.
33. Jackson, R.M., Gabb, H.A. and Sternberg, M.J.E., J. Mol. Biol., 276 (1998) 265.
34. Mendes, J., Baptista, A.M., Carrondo, M.A. and Soares, C.M., Proteins, 37 (1999) 530.
35. van Gunsteren, W.F. and Berendsen, H.J.C. Groningen molecular simulation (GROMOS) library manual, Biomos B. V., Biomolecular Software, Groningen, The Netherlands, 1987.
36. Solmajer, T. and Mehler, E.L., Protein Eng., 4 (1991) 911.
37. Solmajer, T. and Mehler, E.L., Int. J. Quant. Chem., 44 (1992) 291.
38. Smith, L.J., Mark, A.E., Dobson, C.M. and van Gunsteren, W.F., Biochemistry, 34 (1995) 10918.
39. Shrake, A. and Rupley, J.A., J. Mol. Biol., 79 (1973) 351.
40. Wolfenden, R., Andersson, L., Cullis, P.M. and Southgate, C.C.B., Biochemistry, 20 (1981) 849.
41. Ben-Naim, A., J. Phys. Chem., 82 (1978) 792.
42. Sharp, K.A., Nicholls, A., Friedman, R. and Honig, B., Biochemistry, 30 (1991) 9686.
43. Vriend, G., J. Mol. Graph., 8 (1990) 52.
44. Kendall, M. and Stuart, A. The advanced theory of statistics, Charles Griffin & Company Ltd., London, 1977.
45. Privalov, P.L. and Makhatadze, G.I., J. Mol. Biol., 232 (1993) 660.
46. Dill, K.A. and Shortle, D., Ann. Rev. Biochem., 60 (1991) 795.
47. Shortle, D., FASEB J., 10 (1996) 27.
48. Gordon, D.B., Marshall, S.A. and Mayo, S.L., Curr. Opin. Struct. Biol., 9 (1999) 509.
49. Hellinga, H.W. and Richards, F.M., Proc. Natl. Acad. Sci. USA, 91 (1994) 5803.
50. Baldwin, E.P., Hajiseyedjavadi, O., Baase, W.A. and Matthews, B.W., Science, 262 (1993) 1715.
51. Eriksson, A.E., Baase, W.A., Zhang, X.J., Heinz, D.W., Blaber, M., Baldwin, E.P. and Matthews, B.W., Science, 255 (1992) 178.
52. Desjarlais, J.R. and Handel, T.M., J. Mol. Biol., 289 (1999) 305
53. Harbury, P.B., Tidor, B. and Kim, P.S., Proc. Natl. Acad. Sci. USA, 92 (1995) 8408.
54. Koehl, P. and Delarue, M., Nature Struct. Biol., 2 (1995) 163.