Journal of Computer-Aided Molecular Design, **15:** 521–532, 2001.
KLUWER/ESCOM

Use of surface area computations to describe atom–atom interactions

Xavier de la Cruz[a,*] & Miquel Calvo[b]
[a]*Departamento de Bioquímica y Biología Molecular, Facultad de Químicas, Universidad de Barcelona, Martí i Franqués, 1, 08028 Barcelona, Spain;* [b]*Departament d'Estadística, Facultat de Biologia, Diagonal 645, 08028-Barcelona, Spain*

Received 29 June 2000; accepted 23 March 2001

Key words: accessible surface area; atom-atom interactions; protein stability; protein structural analysis; protein structure prediction

Summary

Accessible surface (ASA) and atomic contact (ACA) areas are powerful tools for protein structure analysis. However, their use for analysis purposes could be extended if a relationship between them and protein stability could be found. At present, this is the case only for ASAs, which have been used to assess the contribution of the hydrophobic effect to protein stability. In the present work we study whether there is a relationship between atomic contact areas and the free energy associated to atom-atom interactions. We utilise a model in which the contribution of atomic interactions to protein stability is expressed as a linear function of the accessible surface area buried between atom pairs. We assess the validity of this hypothesis, using a set of 124 lysozyme mutants (Matthews, 1995, Adv Protein Chem, 249–278) for which both the X-ray structure and the experimental stability are known. We tested this assumption for residue representations with increasing numbers of atom types. Our results indicate that for simple residue representations, with only 4 to 5 atom types, there is not a clear linear relationship between stability and buried accessible area. However, this relationship is observed for representations with 6 to 9 atom types, where gross heterogeneities in the atom type definition are eliminated. Finally, we also study a version of the linear model in which the atom- atom interactions are represented utilising a simple function for the buried accessible area, which may be useful for protein structure prediction studies.

The challenge of understanding protein structures, their stability, as well as their complexes, has fuelled the development of different approaches for their study. In this context, surface area computations probably constitute one of the most broadly used tools. They have been used to analyse secondary structure packing [1–2], in studying hydrophobic effect [3–9], etc. However, the usefulness of surface area computations would be further extended if it could be shown that they can also help the researcher in linking more deeply structural and energetic analysis.

At present the best well known relationship linking surface areas and protein stability is the linear relationship between accessible surface area (ASA) and the free energy of transfer of aminoacids from water to organic solvents [3], which is a measure of the hydrophobic effect [5]. However, a series of studies by different authors suggest that surface area computations could also be used to model the contribution, ΔG_{NB}, of atom-atom interactions to protein stability. De la Cruz et al. [10–11] proposed that ΔG_{NB} can be written as:

$$\Delta G_{NB} = \Sigma_{i,j} l_{ij} \cdot \Delta S_{ij} \quad (1)$$

where ΔS_{ij} is the difference between the total atomic accessible surface area buried in the native state between atoms of type i and j, and its value in the denatured state. We call it atomic contact area (ACA). Subindexes i and j vary between 1 and N, with $i \leq j$. N is the number of different atom types used (polar oxygens, aliphatic carbons, etc), and reflects the level of detail of the residue representation used. (A word

*To whom correspondence should be addressed. E-mail: xavier@husky.bq.ub.es

of caution: the symbol S used in Equation (1) must not be confused with that used for entropies in thermodynamics.) The l_{ij} are the proportionality coefficients (Kcal/mol Å^2). The product $l_{ij} \cdot \Delta S_{ij}$ provides an estimate of the total contribution of type i-j atom-atom interactions (which can be hydrogen bonds, carbon-carbon interactions, etc) to protein stability. This product has been shown to be enough to properly reproduce the behaviour of non-bonded interactions [11].

When combined with a typical surface-area-dependent solvation term [9] ΔG_{NB} can be used to estimate the contribution of atomic interactions (both atom-atom and atom-solvent), ΔG_{Int}, to protein stability as follows:

$$\Delta G_{Int} = \Sigma_{i,j}\, l_{ij} \cdot \Delta S_{ij} + \Sigma_i\, l_{iw} \cdot \Delta S_{iw} \qquad (2)$$

where the first term in the right-hand side corresponds to ΔG_{NB}. The second term corresponds to the solvation contribution for which, as for ΔG_{NB}, the sum runs over the different atom types i (polar oxygens, aliphatic carbons, etc) used to model protein residues. l_{iw} is the solvation coefficient for type i atoms. ΔS_{iw} is the total change in accessible surface area between the native state and a given reference state, for type i atoms. The product $l_{iw} \cdot \Delta S_{iw}$ provides an estimate of the contribution to protein stability of the solvation of type i atoms [9].

This model was utilised as a tool for protein structure refinement [10] and analysis [11] with promising results. Koehl and Delarue [12] used a simplified version of Equation (2) in the evaluation of protein models and also in the prediction of free energy changes arising from site-directed mutagenesis experiments. Their results confirmed the usefulness of Equation (2) for the purpose of protein model evaluation. However, its application to the prediction of mutant stabilities produced less clear results, with an excellent correlation (0.90) between experimental and calculated free energy variations for a small number (17) of phage T4 lysozyme mutants, and clearly poorer results (correlation of 0.52) in the case of staphylococcal nuclease mutants. The same authors [13] describe an application of their previous work to the field of fold recognition, utilising a slightly different version of their previous model [12] within the framework of the inverse Boltzmann law.

Different variants of Equation (2) have been used in *ab initio* protein structure prediction studies [14–15] to model the contribution of non-bonded terms to the free energy of the native protein. The results of these authors indicate that Equation (2) can be used to provide a fair first approximation of the effective interaction between hydrophobic residues. An interesting application to the analysis of protein complexes has been recently described [16–17].

In general, the above mentioned results suggest that model (2) could be used to extend surface area computations to provide an approximated understanding of the contribution of atomic interactions to protein stability. This would enlarge the reach of surface area computations and could be of interest in the fields of protein structure prediction [18] and of protein engineering [19, 20] where a quantitative understanding of protein stability is essential. In the present article, we explore the limits and validity of this assumption, using the large amount of structural and thermodynamic data available for Phage T4 lysozyme [21]. The results obtained indicate that the validity of the linear relationship stated in Equation (2) depends on the number of atom types utilised to represent protein residues. Models using a few atom types −4 or 5, and with gross heterogeneities in their definition, yielded low correlation coefficients. A model using 9, more homogeneous, atom types showed the best results. Additionally, as Equation (2) has been used in protein structure prediction studies [14–15] we also tested its performance when the numerically computed ACA's where replaced by a simpler approximation with a lower computational cost.

Materials and methods

(i) The Phage T4 lysozyme mutants

A total of 124 T4 lysozyme replacement mutants were used. They were obtained from a list of mutants provided by Matthews [21], after eliminating those cases for which there were incomplete residues in the determined structure. Both structure and free energy differences, $\Delta\Delta G$, between native and mutants at pH 2 were available for all of them. $\Delta\Delta G$ values varied from −8.9 Kcal/mol to 1 Kcal/mol, with an average of −1.1 Kcal/mol and a standard deviation of 1.61 Kcal/mol.

The coordinates for both the native and the mutant T4 lysozymes were taken from the PDB files listed in Table I from Matthews [21]. Residues 162-164, very mobile both in the native and mutant structures, were eliminated. Following Matthews [21], in some cases the native structure was taken to be the wild type (3lzm PDB code), while in other cases we used the structure of the mutant C54T/C97A, (1l63 PDB code).

Table 1. Results from the statistical tests on the different versions of model (2). (A) results for the numerical ACAs; (B) results for the simplified analytical ACAs ((A) and (B) correspond to the first and second methods to compute the ACAs listed in the Methods).

	r full model[b]	F-test[c]	P.C.A.[d]	r reduced model[e]	# Effect. variables[f]
			(A)		
R6[a]	0.85	9.13	17	0.84	17
R7[a]	0.87	8.13	21	0.86	21
R9[a]	0.90	5.89	28	0.89	28
			(B)		
R6[a]	0.84	8.29	15	0.84	18
R7[a]	0.86	7.23	19	0.85	23
R9[a]	0.88	4.71	26	0.87	28

[a]Residue representation code, see Table III.
[b]Correlation coefficient for the full model, before eliminating any ACA.
[c]F-test for the full model.
[d]Number of effective ACAs, according to the principal component analysis (see Methods).
[e]Same as 2 for the reduced model (see text).
[f]Number of effective variables in the reduced model.

(ii) The free-energy model

The T4 lysozyme mutants were utilised to test equation (2) by using the free energy differences between the mutant and the native protein, $\Delta\Delta G$, provided by Matthews [21] in his Table I. In a few cases the original papers were searched for free-energy values computed at pH closer to 2, as this was the pH for which there were more observations. It has to be noted that the $\Delta\Delta G$ values cited by Matthews [21] do not correspond exactly to the quantity represented by Equation (2), as they also include the contribution of the conformational entropy. This will obviously affect the results, however, as both mutant and native proteins differ in only one or a very small number of residues, it is reasonable to assume that the contribution of the conformational entropy term to $\Delta\Delta G$ will be small. The $\Delta\Delta G$ are expressed as [21]:

$$\Delta\Delta G = \Delta G^N - \Delta G^M \tag{3}$$

where ΔG^N and ΔG^M correspond to free energies of unfolding of the native (N) and the mutant (M) proteins, respectively. To test our approach we modelled them using equation (3), in the assumption that the contribution of conformational entropies can be disregarded. $\Delta\Delta G$ can then be written as:

$$\begin{aligned}\Delta\Delta G = {} & \Sigma_{i,j} l_{ij}.\left(\Delta S^N_{ij} - \Delta S^M_{ij}\right) \\ & + \Sigma_i\, l_{iw}.\left(\Delta S^N_{iw} - \Delta S^M_{iw}\right)\end{aligned} \tag{4}$$

ΔS^N_{iw} is the change in total accessibility for the atoms of type i, between the native and the denatured state, for the native protein. It is computed as:

$$\Delta S^N_{iw} = \sum_m \Delta s^N_{mw} \tag{5}$$

where the index m runs over all the atoms of type i in the native protein. Δs^N_{mw} is equal to:

$$\Delta s^N_{mw} = s^{N,D}_{mw} - s^{N,F}_{mw} \tag{6}$$

where $s^{N,D}_{mw}$ and $s^{N,F}_{mw}$ are the accessible surface area (ASA) of atom m, in the denatured (D) and folded (F) states of the protein, respectively.

ΔS^M_{iw} in (4) is derived in a fashion exactly analogous to ΔS^N_{iw}.

ΔS^N_{ij} in (4) is equal to:

$$\Delta S^N_{ij} = \sum_{m,n} \Delta s^N_{mn} \tag{7}$$

The subindexes m and n, run over all the atoms of type i and j, respectively, in the native protein. Δs^N_{mn} is equal to:

$$\Delta s^N_{mn} = s^{N,D}_{mn} - s^{N,F}_{mn} \tag{8}$$

$s^{N,F}_{mn}$ is the atomic contact area (ACA) between atoms m and n, when the native protein (N) is in the folded state (F). It is computed as described below. $s^{N,D}_{mn}$ is analogous for the denatured state (D).

ΔS^M_{ij} in (4) is derived in a fashion exactly analogous to ΔS^N_{ij}.

The previous computations involve the use of a model for the denatured state of both the native and the mutant proteins. However, it is very difficult to provide a convincing model for the denatured state [22] due to its structural complexity [23]. Usually, the denatured state is modelled using the polypeptide chain in an extended conformation [24]. However, recent studies, both experimental [25] and theoretical [24], have shown that this model constitutes a very poor approximation to the denatured state of a protein, leading to gross overestimates of residue accessibilities.

Therefore, rather than introducing an uncontrolled source of variation, we decided not to model the denatured state structural term, and therefore not to use the $s^{X,D}_{mn}$ (X=M or N).

Then, in the present work, we will approximate $\Delta\Delta G$ using the following expression:

$$\Delta\Delta G \approx \Sigma_{i,j} l_{ij} \cdot \left(\sum_{m',n'} s^{M,F}_{m'n'} - \sum_{m,n} s^{N,F}_{mn} \right) + \Sigma_i l_{iw} \cdot \left(\sum_{m'} s^{M,F}_{m'w} - \sum_{m} s^{N,F}_{mw} \right) \quad (9)$$

Different indexes m,n and m′,n′ have been used to indicate that the number of atoms of each type may be different in the native and in the mutant proteins.

It has to be noted that we will also refer to the ACAs as the regressors, the name given to the independent variables in least-squares fitting [26].

(iii) Surface computations

The ASAs of the different atoms, $s^{M,F}_{m'w}$ or $s^{N,F}_{mw}$, used in the second right-hand side term in (9), were computed using the Lee and Richards, 1971, algorithm.

The ACAs for the different atom pairs, $s^{M,F}_{m'n'}$ or $s^{N,F}_{mn}$ – which we will write as s_{mn} for simplicity, were computed using two different approximations: one based on the use of accessible areas, the second based on a simple analytical function. The first approach is as follows:

- Compute s_{m0}, the ASA of atom m considering only its rigid neighbours – atoms directly bonded to m or atoms being involved in a covalent angle with m.
  - Compute s_{m1}, the ASA of atom m considering its rigid neighbours plus atom n.
  - Repeat the first two steps for atom n, to obtain s_{n0} and s_{n1}. Then s_{mn}, the ACA between atoms m and n, is then defined as:

$$s_{mn} = s_{m0} - s_{m1} + s_{n0} - s_{n1} \quad (10)$$

In the second approach s_{mn} was computed using the following formula [27]:

$$s_{mn} = \pi \cdot \left(\frac{r_m + r_n + 2r_w}{d_{mn}} - 1 \right) \left[(r_m + r_n + 2r_w) d_{mn} - \left(r_n^2 - r_m^2 \right) \right] \quad (11)$$

where r_m, r_n and r_w are the atomic radii of atoms m, n and of water, respectively (r_w was set to 1.4 Å). The atomic radii were taken from Lee and Richards [2]. d_{mn} is the distance between atoms m and n.

In both approximations, s_{mn} was computed only when:

$$d_{mn} < r_m + r_n + 2r_w \quad (12)$$

In addition to this, no s_{mn} was computed between atoms belonging to the same residue or in the cases were one of the atoms in the pair belonged to the main chain of a residue adjacent to the residue including the other atom in the pair.

Table 2. Correlation between the ΔΔG values and the aliphatic carbons accessibilities. The latter are supposed to provide a measure of the magnitude of the hydrophobic effect.

Rep.[a]	R[b]
R4	0.05
R4′	0.05
R4″	0.07
R5	0.07
R6	0.07
R7	0.01
R9	0.01

[a]Name of the different residue representations.
[b]Correlation coefficient.

(iv) The residue representations

We tested the validity of equation (2) for seven different residue representations (Table III). For each of them a given number of atom types were used to describe the 20 natural aminoacids.

The number of ACAs used in equation (2) is a quadratic function of the number of atom types, natyp, in the residue representation:

$$0.5\text{natyp} \cdot (\text{natyp} + 3) \quad (13)$$

However, it may happen that the observed number of ACAs is smaller due to the low abundance of some atom types, e.g. the ACA corresponding to the sulphur-sulphur interaction may be missing for this reason.

(v) Standard statistical procedures

Most of the statistics computations –regression models, Analysis of Variance tables, variable inflation factors, etc- were done using the package SPSS, version 7.5.1.

To compute the set of optimal regressors (Methods (vi)) we used the leaps-and-bounds algorithm [28] as implemented in the SAS package, version 6.12.

(vi) Detection of multicollinearity and number of effective variables

As the residue representations improve, i.e. the number of atom types grows, linear relationships between

Table 3. The seven residue representations used in this work.

Rep.[a]	# Atom types	# ACAs	Atom types[b]
R4	4	14	C,S; m.c.N,O; s.c.p.N,O; s.c. N^+,O^-
R4′	4	14	C,S; N,O; s.c. O^-;s.c. N^+
R4″	4	14	C; O; N; S
R5	5	20	C; N,O; O^-; N^+; S
R6	6	27	C; S; O; O^-; N; N^+
R7	7	35	C; arom.C; S; O; O^-; N; N^+
R9	9	54	C; arom.C; S; O(H); O; O^-; N; arom.N; N^+

[a]Name of the different residue representations.
[b]The different atom types in the representation. m.c.:main chain; s.c.: side chain; p.: polar; N^+: charged N as in Arg and Lys; O^-: charged O as in Glu and Asp; arom. aromatic atom, in Trp, His, Phe and Tyr; O(H): oxygen with a bonded hydrogen atom.

the ACAs started to appear. This effect, the existence of linear relationships between the regressors, is known as multicollinearity [26]. For residue representations showing it, this means that only some of the ACAs are linearly independent and may truly contribute to explain the behaviour of the $\Delta\Delta$Gs. To detect the presence of multicollinearity we followed two strategies: use of a principal components analysis in the regressors [26] and use of the leaps-and-bounds algorithm [28].

The principal components method is based in the diagonalization of the matrix:

$$X'(s) \cdot X(s) \tag{14}$$

where $X_{(s)}$ is the scaled design matrix [26].

This method is based on the fact that the eigenvalue associated with each principal component indicates its contribution to the variability of the ACAs. So, after performing the principal component analysis, the eigenvalues obtained are sorted in decreasing order. Then, starting from the highest value, they are added until the sum reaches a threshold of 95% of the total variability. It is accepted [26] that the number of eigenvalues contributing to this sum is a measure of the number of effectively independent ACAs.

As mentioned before we also used the leaps-and-bounds algorithm [26] to provide an independent measure of the number of effective variables in the model. For a given residue representation, this algorithm explores all the versions of Equation (2) corresponding to all the possible combinations of ACAs until it finds the set which gives the minimal squared error (MSE; see below for its definition). The number of variables in this optimal set constitutes another measure of the number of effective variables in the model. In general, this number will be very close to that provided by the principal components analysis. One advantage of using this method is that it provides the correlation coefficient for the resulting model. We will refer to the optimal models generated by the leaps-and-bounds procedure as reduced models. It has to be noted that the leaps-and-bound algorithm is not a clustering procedure in which variables are grouped into clusters on the basis of a distance matrix. Variables are selected according to their contribution to the predictive power of the model, as measured by the decrease in the MSE of the resulting model after adding the new variable.

The MSE is defined as:

$$\frac{1}{n-p}\sum_{1}^{n}[\Delta\Delta G_i(\text{Obs}) - \Delta\Delta G_i(\text{Pred})]^2 \tag{15}$$

where n is the number of experimental observations available -124 in our case, p is the number of adjustable parameters in equation (9), $\Delta\Delta G_i$(Obs) and $\Delta\Delta G_i$(Pred) are the ith observation and prediction, respectively.

(vii) Assessment of the possibility that a given correlation coefficient may arise by chance

Residue representation R9 was considered of particular interest because several reasons: it showed a good correlation coefficient, $r_{red} = 0.89$, supported by the results of the F-test; the number of atom types, 9, was relatively small; the atom types were the most homogeneous. Then, due to the interest of the model, we decided to provide further evidence confirming the F-test results, to which end we used the following procedures: a jacknife procedure and a permutation procedure.

For the jacknife procedure we followed the standard protocol [29].

The permutation procedure was as follows:

– Generate 1000 sets of $\Delta\Delta$Gs using random permutations of the observed values.
– For each of the 1000 sets find the best set of regressors using the leaps-and-bounds procedure (see (vi)) and obtain the corresponding correlation coefficient.

Then the number of sets for which the correlation coefficient is lower than r_{red}, divided by 1000, is used as a measure of the probability that the value of r_{red} or better correlations were due to chance. This procedure is different from a jacknife procedure in the sense that all the observations are kept in the dataset, which is not the case in the jacknife procedure. In addition, the association between the dependent and the independent variables is altered, that is, after applying the random procedure the 5^{th} observation may now be associated with the 3^{rd} set of regressors, etc.

Results and discussion

To study the validity of model (2) we used several residue representations with an increasing number of atom types, starting with a minimal representation with four atom types. The interest of evaluating small residue representations is twofold: first, they have been used by a number of researchers in the field [9, 12, 13]; second, they allow for a better ratio of observations vs. number of parameters.

Minimal residue representations

For the first residue representation tested we used a set of four atom types, R4, utilised by Delarue and Koehl [13] although in our case the final number of ACAs, 14, was slightly larger. After least-squares fitting of Equation (2) to the experimental data, the correlation coefficient was 0.68. The results of the F-test indicate that this value is significantly different from 0 (p-value < 0.001). However, the relatively low value of the correlation coefficient suggests that with this version of model (2) we can only provide a poor model of the contribution of atom-atom interactions and atom solvation terms to the protein folding free energy. This seems to contradict the results obtained by Delarue and Koehl [13] who found that their model had a good discrimination power when used in fold recognition problems. This apparent contradiction is probably due to the nature of the scoring function they used. Delarue and Koehl [13] utilised an ACA-based statistical potential, derived from database counting and the inverse Boltzmann law, to discriminate between the native structure of a protein and a set of non-native candidates. Their potential reflects structural key features of native proteins which are useful to reject some or all of the non-native structures generated during the threading process. However, Ben-Naim [30] has recently shown that statistical potentials of this kind do not correspond to experimentally measurable interaction free energies. This probably explains why these potentials are of value when applied to the threading problem, while their use is more limited in the computation of protein stability.

To evaluate the effect of the atom type selection, we tested the performance of two additional residue representations, R4′ and R4″ (see Methods), with the same number of ACAs as R4. The values obtained for the correlation coefficients, 0.68 and 0.74 respectively, are very similar to that of R4. This indicates that for this number of atom types, the selection of the residue representation does not lead to substantial improvements in the linear relationship between ΔG and ACAs.

Taken together, these results strongly suggest that very simple residue representations cannot yield a clear linear relationship between ACAs and ΔG.

Better residue representations

One common characteristic of the previous residue models is that their small number of atom types implies that the same atom type will be used for atoms with different properties, e.g. this is the case for N^+ and O^- in the Delarue and Koehl model [13]. This factor limits the possibility to correctly reproduce the interatomic interactions contributing to the protein stability [31]. We therefore decided to improve the residue representations by increasing the number of different atom types and eliminating the most important heterogeneities present in the previous cases (Table III; see Methods).

In the first representation tested we used the five atom types utilised by Eisenberg and McLachlan [9]. The correlation coefficient was still low, 0.73, although significantly different from zero, according to the F-test results (p-value < 0.001). As before, the low correlation coefficient indicates the limited ability of

this version of model (2) to explain the free energy data.

The remaining versions of model (2), derived from the residue representations R6, R7 and R9, had 27, 35 and 54 ACAs, respectively. The corresponding correlation coefficients, 0.85, 0.87 and 0.90, were large and significantly different from 0, according to the F-test results (p-value < 0.001). However, there are two problems with these versions of model (2): the existence of linear relationships between the ACAs, and the large number of ACAs relative to the number of observations. We discuss them below.

(i) Evaluation of the linear relationships between the ACAs

The existence of linear relationships between the regressor variables -the ACAs- is called multicollinearity [32]. In our case multicollinearity arises because of the covalent bonds between atoms of different types, and also probably because of the nature of the mutagenesis experiment, in which residues with atoms of different types are interchanged. One of the consequences of multicollinearity is that the errors in the least-squares estimates of the parameters may be very large, severely limiting the predictive power of the model [26].

We used a principal component analysis (see Methods) to evaluate the extent of the multicollinearity problem in our case. This allowed us to compute the number of effectively independent ACAs for each of the residue representations. The results obtained show (Table IA) that in each case the number of effective (or independent) variables is between 50% and 60% of the total number of variables. To confirm that this number of ACAs is enough give a similar fit to the experimental data as the whole set, an exhaustive search for the optimal set of ACAs was done (see Methods). The results of this standard procedure indicate (Table I) that for each residue representation we could find a reduced version of model (2) in which all the ACAs were independent and with a correlation coefficient very close to that obtained for the full set of ACAs. In all the cases, the number of variables in the reduced model was the same as that found by the principal component analysis.

The results of this section confirm that models R6, R7 and R9 show multicollinearity. This problem strongly affects the quality of the least-squares estimates parameters leading to poor estimates of the experimental values. However, it does not prevent the assessment of the applicability of equation (2).

(ii) The large number of ACAs in the R6, R7 and R9 representations

As mentioned before, the results of the F-tests indicate that the correlation coefficients for R6, R7 and R9 are significantly different from zero. However the fact that R6, R7 and R9 are associated with a high number of ACAs suggests that there is a chance that the observed improvements in the correlation coefficients may have a spurious origin. In this respect it has to be mentioned that the results of the multicollinearity analysis show that the effective number of ACAs is smaller than their total number: 17 vs. 27 for R6, 21 vs. 35 for R7, and 28 vs. 54 for R9. These numbers show that the effective number of ACAs contributing to explain the ΔG variability is clearly lower, thus giving a better ratio between the number of observations and that of variables in the model. In combination with the results of the F-tests, the fact that the new atom types are more chemically meaningful, these results provide stronger evidence to the likelihood that the observed linear relationships are not due to a pure chance effect.

It has to be noted that the correlation coefficients observed for R6, R7 and R9 are not trivially due to changes in the accessibility of the aliphatic carbons, known to relate linearly to free energy changes [6–8]. Actually, the individual correlation coefficients for the ACAs related to the hydrophobic effect are low (Table II). This is in accordance with the fact that the lysozyme mutants were obtained to study a broad set of different contributions to the free energy of folding [21].

(iii) The R9 representation

As seen from the above results, R9 is the residue representation giving the highest correlation coefficient for both the full and the reduced models. In addition, in both cases the residual followed a clear Normal distribution (results not shown), as expected in well behaving least-squares models [24]. Also the results of the F-test for the full and the reduced models showed that the correlation coefficient is different from zero with a p-value lower than 0.001. This suggests that R9 may constitute a minimal set of atom types useful for protein stability studies.

The validation of the results obtained for R9 was done applying the jacknife procedure [29] to our problem. The jacknife estimate of the correlation coeffi-

cient [33] was 0.80. As expected, this value is slightly lower than the all data estimate, 0.89. However, it is high enough to support the existence of the linear relationship described in Equation (2).

Finally, to compliment the results of the jacknife procedure, we applied to our data a permutation procedure as described in Methods (vii). This approach, despite some limitations [34], has the virtue of providing a simple and intuitive measure of the likelihood that the correlation observed for R9 would appear by chance. Our results show that for none of the 1000 random samples generated from the $\Delta\Delta$Gs the correlation coefficient was higher than that observed for the original dataset. In addition, the z-score of the MSE of the reduced model is of - 4.2 while that of the closest random model is of -3.4. These results suggest that for R9 there is a low probability that the linear relationship observed for the reduced model is due to chance factors.

The predictive power of the method is at present limited by the low quality of the l_{ij} estimates which is due to:

- The size of the experimental error in the observed $\Delta\Delta$G values. Relative errors are bigger than 50% in more than 30% of the cases, introducing a substantial error in the l_{ij} estimates [26].
- The existence of multicollinearity in the data [26].

However, in spite of these limiting problems, the use of ΔS_{ij} provides an intuitive understanding for many of the stability changes due to mutations. For example, to study the contribution of hydrogen bonds to protein stability, Matthews and coworkers [35] built a series of 13 mutants at position 157 of lysozyme T4. At that position, threonine, the wild type residue, is involved in a hydrogen bond net through its hydroxyl group. Between the 13 mutants, only T157S preserved this hydrogen bond net. On the contrary, the non-disruptive mutant T157L lost all the native hydrogen bonds involving residue 157 side-chain hydroxyl group. If we compare both mutants we can see that there a decrease of 50 $Å^2$ from T157S to T157L in the ΔS_{ij} corresponding to the hydrogen bond interaction. This difference is due to the loss of the native hydrogen bonds in T157L and corresponds to a computed free energy difference of 0.25 Kcal/mol between both mutants. Although this value is lower than the observed stability change between the mutants, 0.5 Kcal/mol, it points in the right direction. Similar results are obtained for the remaining mutants, although for the more disruptive mutations the hydrogen bond term was not enough to explain the free energy differences.

This example shows that the use of ΔS_{ij} allows a good qualitative understanding of protein stability changes due to mutations, by pointing to relevant changes in the atomic interactions. However, better l_{ij} are required to obtain a more quantitative understanding. A possible approach to derive them would be the use of the Wallqvist et al. method [16] which utilise ACA-based statistical potentials to model atom-atom interactions. However, despite the undeniable value of this approach, these parameters may not correspond to the measurable interaction free energies [30]. The best option is, a priori, to improve parameter quality through an increase in the mutants dataset.

The previous comments bring the problem of the transferability of the parameters, that is, whether the parameters derived from one set of mutants in one protein can be used to study another set of mutants in a different protein. Obviously, if our model for atom-atom interactions were exact and we were able to include all the remaining contributions to $\Delta\Delta$G transferability would not be a problem, as the physical basis for atom-atom interactions is the same in one or another protein. However, the lack of some terms –e.g. conformational entropy, secondary structure propensities, covalent terms- and departures from linearity contribute to limit the transferability of these parameters. This is so because the fitting procedure will tend to biass the adjusted parameters in order to produce a better fit to the observations. As some of the missing properties will change from one protein to another, or from one set of mutants to another, the biass will change correspondingly from one case to another. Unfortunately, the limited number of data available at present does not allow a proper modelling of these missing terms, thus limiting to some extent the transferability of the derived parameters.

An interesting case arises when considering the work by Takano et al., 1998 [36]. The authors build a series of mutants Ile–Val and Val–Ala in order to study the contribution of the hydrophobic effect to the resulting stability changes. When clustering their mutations according to the secondary structure of the mutated residue Takano et al. [36] find, in general, a good correlation between $\Delta\Delta$G values and changes in the amount of exposed hydrophobic surface area. This is true for the five Val–Ala mutations located in helices. Interestingly, because these mutations happen at mostly buried locations, this result would be in apparent contradiction with the theory of protein engi-

neering analysis of stability by Fersht and coworkers [37]. In this theory free energy changes caused by non-disruptive mutations, like Val–Ala, in buried groups can be written as [37]:

$$\begin{aligned}\Delta\Delta G = \; & G_{F(X..Y)} + G_{F(X..E)} \\ & - G_{U(X..Water)} + G'_{U(H..Water)}\end{aligned} \tag{16}$$

where X is the group lost in the mutation, H is the replacing atom, Y represents a specific group interacting with X in the native protein and E corresponds to the remaining groups interacting with X. $G'_{U(H..Water)}$ and $G_{U(X..Water)}$ are the solvation terms of H and group X in the unfolded state, for the mutant and native proteins, respectively. $G_{F(X..Y)}$ and $G_{F(X..E)}$ are the terms which account for the atom-atom interactions lost in the native state when deleting group X. Because these two terms are essentially analogous, we can rewrite (16) as:

$$\Delta\Delta G = G_{F(X..C)} - G_{U(X..Water)} + G'_{U(H..Water)} \tag{17}$$

where $G_{F(X..C)} = G_{F(X..Y)} + G_{F(X..E)}$ is the contribution of the atom-atom interactions involving group X. $G_{F(X..C)}$ has the same meaning as our ΔG_{NB} term (which we defined in the introducion as the contribution of all atom-atom interactions to protein stability).

The same $\Delta\Delta G$ change is written by Takano et al., as [36]:

$$\begin{aligned}\Delta\Delta G = \; & a \cdot (ASA_{fold,wild} - ASA_{fold,mut} \\ & - ASA_{unf,wild} + ASA_{unf,mut}) + b\end{aligned} \tag{18}$$

where $ASA_{fold,wild}$ and $ASA_{fold,mut}$ correspond to the hydrophobic surface area exposed in the native state of the wild and the mutant proteins, respectively; and $ASA_{unf,wild}$ and $ASA_{unf,mut}$ correspond to the hydrophobic surface area exposed in the denatured state of the wild and the mutant proteins. Finally a and b are the adjustable parameters of the model.

If we assume a linear relationship between exposed atomic surface area and solvation free energy [3, 9] a reasonable equivalence can be stablished between $G_{U(X..Water)}$ and $a.ASA_{unf,wild}$, as well as between $G'_{U(H..Water)}$ and $a.ASA_{unf,mut}$. This leaves us with the following relationship:

$$G_{F(X..C)} \approx a \cdot \Delta ASA + b \tag{19}$$

where $\Delta ASA = ASA_{fold,wild} - ASA_{fold,mut}$.

The apparent contradiction arises because the left-hand side term of the equation may comprise a set of heterogeneous atom-atom interactions, while the right-hand side term only takes into account changes in the hydrophobic surface area exposed to the solvent, plus a constant term. However, the contradiction can be eliminated if we assume:

– That there is a linear, or approximately linear, relationship between ΔASA and S_{Xj}, the total ACA corresponding to the interaction between group X and j-type atoms belonging to the X group environment:
$$S_{Xj} \approx a_{Xj}.\Delta ASA + b_{Xj} \tag{20}$$
where $\Delta ASA = ASA_{fold,wild} - ASA_{fold,mut}$. For simplicity we assume that group X is constituted by atoms of the same type. This is clearly the case in Val–Ala mutations. Considering more than one atom type would require only small changes in the equations but would not affect the final conclusions.
– The validity of our model. That is, that there is a linear relationship, as in equation (1), between $G_{F(X..C)}$ and the S_{Xj} describing the environment of group X:
$$G_F(X..C) = \Sigma_{Xj} l_{Xj}.S_{Xj} \tag{21}$$
where subindex j in the right-hand side sum runs over all atom types belonging to X group environment.

Then replacing S_{Xj} in (21) by its value in (20) we find that:

$$G_{F(X..C)} \approx \Sigma_{Xj} l_{Xj}.(a_{Xj}.\Delta ASA + b_{Xj}) \tag{22}$$

Which can be rewritten as:

$$G_{F(X..C)} \approx (\Sigma_{Xj} l_{Xj}.a_{Xj})\Delta ASA + \Sigma_{Xj} l_{Xj}.b_{Xj} \tag{23}$$

We can see that equation (23) is formally equivalent to (19).

We first look at the assumption on the existence of a possible linear relationship between the S_{Xj} and the ΔASA in the case of the Val–Ala mutants. Only interactions involving aliphatic carbons are considered, as these are the more likely to show substantial changes in the Val–Ala mutations. For the nine possible interactions between aliphatic carbons and the remaining atom types – when using representation R9, the computed correlation coefficients between the corresponding S_{Xj} and the ΔASA varied between -0.07 and 0.90. It has to be noted that for those interactions involving larger amounts of S_{Xj} the correlation coefficients were the highest. For example, the largest value, 0.90, was observed for the correlation between ΔASA and the aliphatic-aliphatic S_{Xj}, which accounts for 29% of the total S_{Xj}. In addition, an average correlation coefficient of 0.73 was observed for those S_{Xj} accounting for 90% of the total of the S_{Xj}. Low correlation coefficients were only observed for a few less

populated interactions. This is probably due to the fact that the latter are more sensitive to changes induced by the mutations in the local structure around the mutation site.

The previous results confirm that for the Val–Ala mutants described there is an approximate linear relationship between S_{Xj} and the ΔASA. Therefore, assuming the validity of our model, a simple explanation can be provided to reconcile Takano et al. [36] data with the theory of Fersht et al. [37]. Although these results have to be taken with care, because of the small number of Val–Ala mutations, they illustrate the explanatory power of our model.

Possible origins of the non-linear term in ΔG

Our results are consistent with the existence of a component in ΔG that has a linear dependence on the ACAs, as expressed in (2). However, there is also a non-linear component in the ΔG, which cannot be modelled using equation (2). Apart from the problems due to the use of a small number of atom types, there may be different contributions to the departures from linearity. One possible source of non-linear behaviour can be the fact that the denatured state was not taken into account when fitting the model to the $\Delta\Delta G$ data. This approximation will affect more directly the l_{iw} used to model the solvation contribution to $\Delta\Delta G$, because denatured states have bigger amounts of exposed surface area than the native protein. However, we believe that the error due to this cause is likely to be small because we are using $\Delta\Delta$Gs instead of ΔGs and it is probable that the differences between the native and the mutant denatured states are not very big.

Conformational entropy is not explicitly modelled in equation (2). This approximation may a priori have an effect on the predictive ability of the model. However, this effect is probably small as Abagyan and Totrov have shown [38] that accessible surface areas can be used to model the contributions to protein stability arising from side-chain conformational entropy. This means that the l_{ij} parameters obtained after the fit to experimental data may include the side-chain contribution to $\Delta\Delta G$.

The main-chain contribution to the conformational entropy can be associated with changes in the backbone flexibility due to the mutations [39]. A raw estimate of these changes can be derived from crystallographic B-factors [40]. In Figure 1 we show a histogram of the difference between the averaged

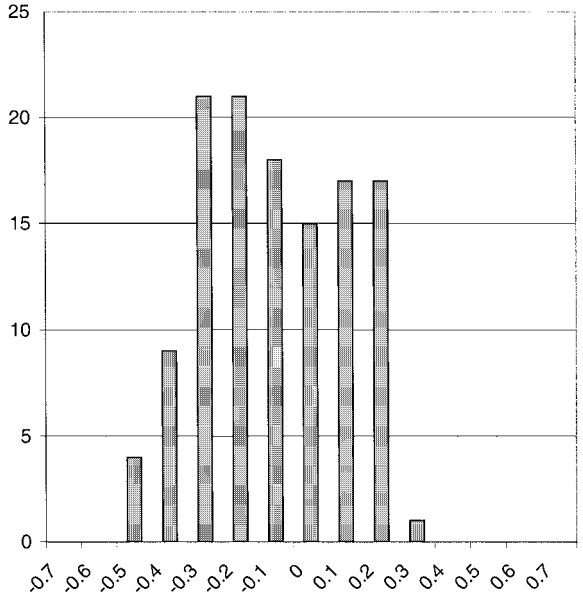

Figure 1. Histogram of the difference in average atomic fluctuations between native and mutant proteins, for main-chain atoms. Abscissae correspond to the atomic fluctuations in Å. These were computed using the B-factors available from the X-ray data [40] for each protein.

main-chain atoms fluctuations for the mutant and the native. We can see that there is a small biass towards negative values indicating that mutant proteins in the native state may have a slightly higher conformational entropy than the wild-type, when considering backbone atoms. The true contribution of this effect is overestimated in this simple approach because isotropic B-factors do not take into account collective motions and because of the contribution of lattice disorder to B-factors [40]. In addition, to properly assess the contribution of this term we would also need to know the contribution of the denatured states. This second term cannot be evaluated from the data we have at hand. It seems reasonable to assume that a similar biass may appear when considering main-chain flexibility for the unfolded states. If this were the case the folded and unfolded contributions could cancel and the effect of this term on the adjusted parameters would be small. We then believe that, at the level of approximation used in assessing backbone flexibility from B-factor values, neglecting the main-chain conformational entropy term is not likely to have introduced a major biass in the l_{ij} values.

Other possible sources of non-linear behaviour are:

– The complex dependence of the values of the ACAs on the interatomic distance, due to the

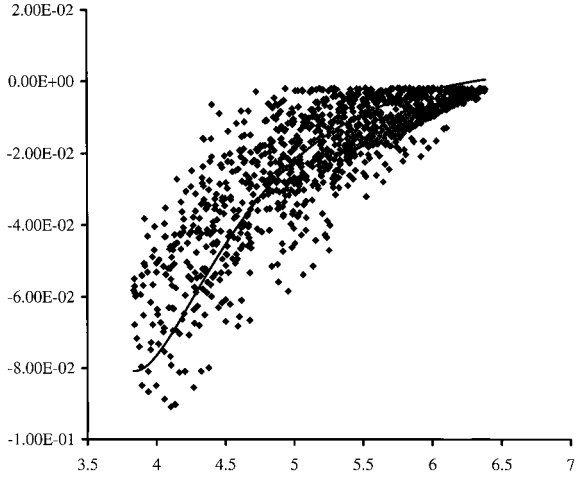

Figure 2. Distance dependence of the van der Waals + Electrostatic (continuous line) and the ACA-based curves (x symbol) for the carbon-carbon interaction. The parameters for the former correspond to the CT-CT interaction in the AMBER force field [48], where CT is one of the carbon atom types in AMBER. The ACA-based curve is a simple proportional relationship of the kind: $l.S_{CT,CT}$, where l is the proportionality constant and $s_{CT,CT}$ is the numerically computed ACA between CT carbon atoms pairs. The coordinates for the latter were obtained from the T4 lysozyme PDB file, 3lzm. The proportionality constant l was obtained after fitting the ACA-based curve to the AMBER CT-CT interaction curve.

presence of the covalent neighbours. This translates into a poorer ability to reproduce the curves representing the van der Waals and electrostatic interactions (Figure 2).
- The energy terms not included in the model, e.g. changes in the covalent bonds and angles, secondary structure propensities, etc.
- The limits of the surface area proportionality law to reproduce the free energy of solvation of polar groups [22, 41].
- The breakage of the additivity hypothesis which states that the different contributions to the free energy are additive [42–43].
- The use of accessible surface areas instead of molecular surfaces [44–45].

It is very difficult to evaluate the contribution of the previous effects to the non-linearity of the free energy. However, it is likely that in our case measurement errors in the $\Delta\Delta$Gs will overshadow this other effects. Actually, in the cases for which an estimate is given, experimental errors may vary between 10% and 90% of the $\Delta\Delta$G values [46, 47]. As mentioned before, these high relative errors can clearly affect the quality of the fit as well as the magnitude of the observed correlation coefficients.

A simple atom-atom interaction model

The previous results indicate that the contribution of atom-atom interactions to the free energy of protein folding can be modelled, in a first stage, using a simple linear relationship in the ACAs. This simple relationship could be utilised in protein structure prediction studies to replace the more conventional distance dependent non-bonded terms. This would require the computation of the ACAs at each step of the minimization process. Unfortunately, the ACAs are computed numerically using an algorithm which is not fast enough for structure prediction studies. We then decided to see whether a simpler measure of the ACAs could be used in model (2) without affecting too much the observed linear relationship. To this end, for each interacting atom pair we computed the corresponding ACA independently from the remaining protein atoms, using a simple analytical function (see Methods (iii)). As mentioned before, this measure of the contact area [11] allows the reproduction of atom-atom interactions like the van der Waals and electrostatic terms. It has also been shown [14] that it may be utilised to provide a fair approximation of the hydrophobic contribution to the free energy of protein folding.

We tested this approach for R6, R7 and R9. The correlation coefficients obtained were close to those described for the numerically computed ACAs (Table IB). The use of the principal component analysis also showed the existence of a strong multicollinearity effect for each representation. Subsequent use of the exhaustive search procedure (see Methods) confirmed that the observed correlations could also be obtained for a smaller set of independent variables. These results show that, despite the presence of non-linear effects, the simplified ACAs could be used to model the linear component present in protein folding free energies. Therefore, they could be utilised to produce raw models of the native conformation of proteins in prediction studies. This would be in accordance with recent results by Kurochkina and Lee [14] who used a similar approach to compute the ACAs and showed that it could be utilised to successfully recognise the native structure of the ROP dimer from a small set of decoys.

Conclusions

We have assessed whether ACAs can be utilised to evaluate the contribution of atom-atom ineractions to

protein stability, using a set of lysozyme mutants obtained by Matthews et al. [21]. Our results indicate that, using residue representations for which atom types are relatively homogeneous, there is a linear component in ΔG which can be modelled using equation (2). This also indicates that ACAs are not only a powerful structure analysis tool, but they can also provide a link from protein structure to protein stability. Unfortunately, the existence of multicollinearity between the ACAs introduces large errors in the least-squares estimates of the parameters in the model, limiting for the moment the predictive power of this approach. More experimental data would be required to properly overcome this problem.

Finally, our results show that a simpler version of the ACAs can be used in Equation (2), preserving most of the linear behaviour observed for the numerically computed ACAs. This result suggests that these simple ACAs could be useful in protein structure prediction studies.

Acknowlegements

The authors wish to acknowledge Dr I. Fita for his support; Dr R.M. Jackson for useful comments on the manuscript, and the referees for their suggestions. Both authors acknowledge experimentalists for making their results available.

References

1. Richards, F.M., Ann. Rev. Biophys. Bioeng. 6 (1977) 151.
2. Lee, B. and Richards, F.M., J. Mol. Biol. 55 (1971) 379.
3. Chothia, C., Nature 248 (1974) 338.
4. Richards, FM., In Creighton, TM (Ed.) Protein Folding, Freeman and Company, 1992, pp. 1-58.
5. Dill, K.A., Biochemistry 29 (1990) 7133.
6. Matsumura, M., Becktel, W.J. and Matthews, BW., Nature 334 (1988) 406.
7. Serrano, L. and Fersht, A.R., Curr. Opin. Struc. Biol. 3 (1993) 75.
8. Takano, K., Yamagata, Y. and Yutani, K. A., J. Mol. Biol. 280 (1998) 749.
9. Eisenberg, D. and McLachlan, A.D., Nature 319 (1986) 199.
10. de la Cruz, X. and Fita, I., J. Appl. Crys. 24 (1991) 941.
11. de la Cruz, X. , Reverter, J, Fita, I., J. Mol. Graphics 10 (1992) 96.
12. Koehl, P. and Delarue, M., Proteins: Struc. Func. Genet. 20 (1994) 264.
13. Delarue, M. and Koehl, P., J. Mol. Biol. 249 (1995) 675.
14. Kurochkina, N. and Lee, B., Protein Eng. 8 (1995) 437.
15. Yue, K. and Dill, K.A., Protein Sci. 5 (1996) 254.
16. Wallqvist, A., Jernigan, R.L. and Covell, D.G., Protein Sci. 4 (1995) 1881.
17. Covell, D.G. and Wallqvist, A., J. Mol. Biol. 269 (1997) 281.
18. Dill, K.A., Curr. Opin. Struct. Biol. 3 (1993) 99.
19. Vasmatzis, G. and Lee, B. Curr. Opin. Biotech. 8 (1997) 423.
20. Vogt, G., Woell, S. and Argos, P., J. Mol. Biol. 269 (1997) 631.
21. Matthews, B.W., Adv. Protein Chem. 46 (1995) 249.
22. Lazaridis, T., Archontis, G., Karplus, M., Adv. Protein Chem. 47 (1995) 231.
23. Dill, K.A and Shortle, D., Annu. Rev. Biochem. 60 (1991) 795.
24. Creamer, T.P., Srinivasan, R. and Rose, G.D., Biochemistry 34 (1995) 16245.
25. Murphy, L.R., Matubayasi, N., Payne, V.A. and Levy, R.M., Fold.Des. 3 (1998) 105.
26. Sen, A. and Srivastasa, M., Regression Analysis. Theory, Methods and Applications, Springer_Verlag, New York, 1990.
27. Wodak, S. and Janin, J. Proc. Natl. Acad. Sci. USA 77 (1980) 1736.
28. Furnival, G.M. and Wilson, R.B., Technometrics 16 (1974) 499.
29. Efron, B. And Tibshirani, R.J., An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.
30. Ben-Naim, A., J. Chem. Phys. 107 (1997) 3698.
31. Zhang, C., Vasmatzis, G., Cornette, J.L. and DeLisi, C., J. Mol. Biol. 267 (1997) 707.
32. Rawlings, J.O., Applied Regression Analysis, Wadswoth & Brooks/Cole, 1988.
33. Zheng, B. and Agresti, A., Statis.Med. 19 (2000) 1771.
34. Tibshirani, R. and Knight, K., J.R.Statis.Soc 61 (1999) 529.
35. Alber, T., Dao-pin, S., Wilson, K., Wozniak, J.A., Cook, S.P. and Matthews, B.W., Nature 330 (1987) 41.
36. Takano, K., Yamagata, Y. and Yutani, K., J.Mol.Biol. 280 (1998) 749.
37. Fersht, A., Matouschek, A. and Serrano, L., J.Mol.Biol. 224 (1992) 771.
38. Abagyan, R. and Totrov, M., J.Mol.Biol. 235 (1994) 983.
39. Matthews, B.W., Nicholson, H. and Becktel, W.J., Proc.Natl.Acad.Sci.USA, 84 (1987) 6663.
40. Petsko, G.A. and Ringe, D., Annu.Rev.Biophys.Bioeng. 13 (1984) 331.
41. Lazaridis, T. and Karplus, M., J. Mol. Biol. 288 (1999) 477.
42. Mark, A.E. and van Gunsteren, W.F., J. Mol. Biol. 240 (1994)167.
43. Dill, K.A., J. Biol. Chem. 272 (1997) 701.
44. Tuñon, I., Silla, E., Pascual-Ahuir, J.L., Protein Eng. 5 (1992) 715.
45. Jackson, R.M. and Sternberg, M.J.E., Nature 366 (1993) 638.
46. Nicholson, H., Anderson, D.E., Dao-pin, S. and Matthews, B.W., Biochemistry 30 (1991) 9816.
47. Heinz, D.W., Baase, W.A. and Matthews, B.W, Proc. Natl. Acad. Sci. USA 89 (1992) 3751.
48. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. and Kollman, P.A., J. Am. Chem. Soc. 117 (1995) 5179.